Facultad de Ciencias

Aplicación de técnicas de Machine Learning al estudio del cáncer de endometrio

Trabajo de Fin de Máster
para acceder al

MÁSTER EN CIENCIA DE DATOS

Autor: Juan Shallcrass Susinos
Director: Francisco Matorras Weinig
Director: Javier Tomás Anchuelo Latorre

Septiembre - 2020

Resumen

En este trabajo, se estudiará la influencia de distintas variables en el exitus de pacientes operadas de cáncer de endometrio. Para ello, se analizarán dos grupos de variables, uno que contiene aquellas variables que dan información sobre la paciente y el tumor, y otro que contiene variables que dan información sobre el tratamiento aplicado. Con este fin, se utilizará un modelo mixto combinando técnicas de Machine Learning basadas en árboles de clasificación y técnicas de estadística clásica para entrar más en detalle en la influencia de las variables. Con las técnicas de Machine Learning, se definirán dos grupos de pacientes a partir de una división en la variable Estadio FIGO, donde uno de los grupos tiene un diagnóstico a priori positivo mientras que el otro tiene un diagnóstico inicial negativo. Finalmente, se estudiará para los dos grupos qué variables influyen en el exitus de la paciente utilizando el Test de Fisher, y si estas influyen aumentando o disminuyendo las probabilidades de supervivencia, empleando para ello el intervalo de confianza binomial según la modificación del intervalo de Clopper-Pearson propuesta por Lancaster.

Palabras clave: cáncer de endometrio, estadio FIGO, árbol de clasificación, test de Fisher, intervalo de confianza binomial.

Abstract

In the present work, the influence of several variables on the exitus of patients operated on for endometrial cancer will be studied. In order to do so, two groups of variables will be analyzed, one containing those variables that provide information about the patient and the tumor, and another one including those that inform about the treatment applied to the patient. For this purpose, a mixed model will be used combining Machine Learning techniques based on classification trees and classical statistics techniques to analyze more in detail the influence of the variables. With the Machine Learning techniques, two groups of patients will be defined from a division in the variable Estadio FIGO, where one of the groups has a positive initial diagnosis while the other one has a negative initial diagnosis. Finally, Fisher's test will be used to study which variables influence the most on the patient's exitus for both these groups, and the binomial confidence interval as the modification of the Clopper-Pearson interval proposed by Lancaster will be employed to understand whether these variables increase or decrease the probability of survival.

Key words: endometrial cancer, FIGO staging, classification trees, Fisher's test, binomial confidence interval.

Contenidos

1. Introducción

El endometrio es una capa mucosa que recubre el interior del útero. Esta capa es la que se derrama en cada menstruación, y se renueva en cada ciclo menstrual en caso de no haber fecundación. El cáncer de endometrio es el cáncer ginecológico más común en los países desarrollados. Este trabajo se centrará en estudiar este tipo de cáncer en pacientes operadas y analizar qué características afectan más a su gravedad y qué factores pueden mejorar su pronóstico.

Para ello, se cuenta con una serie de datos sobre las pacientes, sus resultados oncológicos, los tratamientos aplicados contra el tumor y posibles recaídas y los lugares en los que se produjeron. Todas las pacientes de las que se tienen datos fueron tratadas en el Hospital Universitario Marqués de Valdecilla, y se obtuvo su consentimiento informado escrito. Además, se pidieron los números de historia clínica al Servicio de Documentación y con ello, permiso al hospital para poder revisar las historias. Para el análisis, estos datos se dividieron en tres categorías según el momento en el que se dispone de la información y el control que se tiene sobre ella.

Así pues, la primera categoría de los datos correspondería con los datos personales de la paciente y las características del tumor, como pueden ser la edad de la paciente, el tamaño o la histología del tumor. Estos datos se caracterizan porque son características no controlables conocidas en un diagnóstico inicial, ya sea de la paciente o del tumor. La segunda categoría de los datos se corresponde con aquellos datos del tratamiento oncológico que son controlables por el médico responsable de la operación. En esta categoría entrarían variables como el tipo de cirugía que se aplica a la paciente, la dosis de radioterapia que se le administró, o si recibió o no quimioterapia. La última categoría que se determinó corresponde a aquellas variables que se conocen después de haber aplicado el tratamiento a la paciente, como pueden ser las toxicidades que se producen a consecuencia del tratamiento, si la paciente ha tenido o no recaída, o los tratamientos de rescate en caso de que haya habido algún tipo de recaída.

En una primera instancia, la intención era la de entrenar una función utilizando únicamente técnicas de Machine Learning que permitiera predecir el exitus de la paciente en función de las variables de las que se dispone. Una vez entrenada esta función, la idea era utilizarla para analizar la dependencia del exitus con estas variables, modificando estas variables para analizar cómo variaba el exitus de la paciente. Sin embargo, debido a la reducida cantidad de datos con los que se contaba y la naturaleza de los mismos, se decidió optar por un modelo mixto, en el cual se mezclan técnicas de Machine Learning para el entrenamiento con técnicas de estadística clásica para el análisis de los resultados.

El trabajo se va a centrar principalmente en los dos primeros grupos de variables. Con el primer grupo, se intentará identificar cuáles son las variables que tienen una mayor influencia en el exitus de la paciente, tratando así de dividir a las pacientes según el diagnóstico inicial para aplicarles un determinado tratamiento. Una vez hecha esta división, se estudiarán qué variables del segundo grupo pueden mejorar el exitus de la paciente, ya sea evitando usar algún determinado tratamiento o utilizando especialmente alguno de ellos. Es importante destacar que el proceso de entrenamiento de

los algoritmos ha sido un proceso “ciego”, en el cual no se ha aportado nunca ningún conocimiento médico adicional. Por lo tanto, el objetivo final del trabajo será tratar de identificar qué variables controlables por el médico pueden ser más efectivas a la hora de tratar a las pacientes en función de su diagnóstico inicial, tratando de analizar la viabilidad de los métodos más que resultados médicos concretos.

2. Métodos y Descripción de los datos

En primer lugar, el lenguaje de programación que se empleó a lo largo de todo el trabajo fue Python. Python es uno de los lenguajes más utilizados en el mundo de la programación [1], principalmente debido a su sencillez de uso y a la gran variedad de posibilidades que ofrece. Debido al alto número de usuarios que lo utilizan, Python ofrece un considerable surtido de librerías que permiten realizar todo tipo de operaciones. Esta es otra de las razones por las que se eligió este lenguaje, ya que tiene disponibles todas las librerías necesarias para el desarrollo del proyecto.

Una de las librerías más usadas y más útiles de este lenguaje es *Pandas* [2]. Esta es una librería especialmente apropiada para cargar, explorar y procesar datos que vienen en un formato tabular. En esta ocasión, los datos fueron proporcionados en un fichero Excel, por lo que se determinó que *Pandas* era la librería adecuada para explorar y tratar los datos dentro de Python [3].

En cuanto a los datos en sí, se cuentan con más de 60 columnas que contienen información sobre la paciente, como su edad, el tumor, su tamaño y otros datos oncológicos como el grado o el estadio patológico, que es una forma de describir cuánto cáncer hay en el cuerpo y en qué partes está localizado, y los tratamientos utilizados, por ejemplo el tipo de radioterapia que recibió la paciente, si ha recibido o no quimioterapia o si hubo recaída local del tumor. La mayoría de estas variables son variables categóricas, y casi todas están categorizadas como números. En algunos casos, estas categorías son progresivas, siendo la categoría más baja la menos grave y la más alta la que tiene más gravedad, mientras que en otros casos, los grupos que se crean no tienen una relación de continuidad entre sí. Para el análisis, se cuenta con los datos de unas 340 pacientes. Este conjunto de datos es un poco atípico para un proyecto de Data Science, ya que se dispone de una cantidad bastante reducida de datos. Este es el principal motivo por el que se optó por utilizar un modelo mixto en lugar de solamente técnicas de Machine Learning, ya que con tan pocos datos es más difícil aplicar este tipo de técnicas.

En un primer lugar, se había planteado la opción de utilizar redes neuronales. Tanto las redes neuronales como los algoritmos basados en los árboles de decisión predicen la variable objetivo, que en este trabajo es el exitus de la paciente, basándose en las variables predictoras que se utilizan para su entrenamiento. En ambos métodos, se puede estudiar cómo afectan las variables a la decisión final, y es por eso por lo que se consideraron los métodos más adecuados para el trabajo. Sin embargo, finalmente se usaron los métodos basados en los árboles de decisión principalmente por el conjunto de datos con el que se contaba. Como se ha mencionado previamente, la mayoría de las variables estaban categorizadas, y las redes neuronales no trabajan del todo bien con este tipo de variables. Además, los árboles de decisión son muy sencillos de entender y de visualizar, ya que se puede ver de una manera muy rápida qué variables son más importantes y qué decisiones se toman para la predicción del exitus.

Para el análisis de los datos, se optó por usar un modelo conocido como árbol de decisión. El árbol de decisión es un modelo de predicción de aprendizaje supervisado que se puede utilizar tanto para clasificación como para regresión. El objetivo de un árbol de decisión es predecir la variable

objetivo mediante el aprendizaje de una serie de reglas basadas en las variables introducidas para su entrenamiento. Para ello, el árbol va dividiendo el conjunto de datos, de tal forma que en las hojas finales del árbol se le asigna un determinado valor al conjunto de datos resultante en cada hoja. Para determinar qué división hacer al conjunto de datos, el árbol utiliza un coeficiente que determina la heterogeneidad del conjunto de datos. Por lo tanto, la división que más reduzca esta heterogeneidad será la división que realice el árbol.

Además del árbol de decisión único, en el trabajo se utilizaron dos modelos más basados en los árboles de decisión y en el aprendizaje por ensemble. El aprendizaje por ensemble se basa en la idea de generar varios modelos débiles y combinarles para generar un modelo fuerte que mejore la estabilidad y el rendimiento del modelo individual. Este tipo de enfoque es especialmente interesante en los árboles de decisión, ya que puede reducir la inestabilidad de los árboles de decisión y mejorar la predicción. El primero de ellos se conoce como Random Forest, y está basado en la técnica de bagging [4]. El bagging obtiene un número determinado de sub-muestras de la muestra de entrenamiento, y crea un árbol individual para cada una de las sub-muestras. Para tomar la decisión final, se basa en la "votación de la mayoría", es decir, en la predicción mayoritaria de cada uno de los árboles individuales. En la Fig. 1 se puede ver un ejemplo de este algoritmo. En este ejemplo, se crean 9 árboles individuales a partir de 9 sub-muestras del conjunto principal de los datos. Cada uno de ellos hace una predicción, y la decisión final es aquella que se haya repetido más veces en todos los árboles.

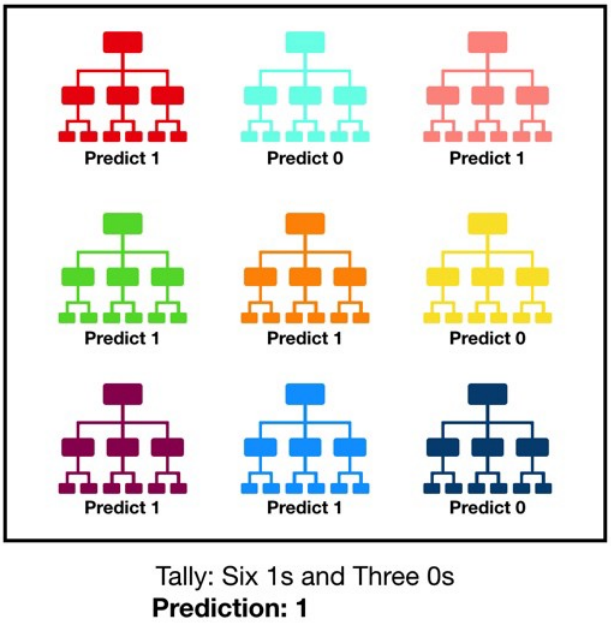

Figura 1: Representación gráfica del algoritmo de Random Forest.

El otro de los modelos es el Gradient Boosting [5]. En el Gradient Boosting están implicados tres elementos: una función de pérdida a optimizar, un modelo débil para hacer predicciones (un árbol de

clasificación en este caso) y un modelo aditivo, que va sumando los modelos débiles para minimizar la función de pérdida. Así pues, en cada una de las iteraciones del proceso, se va añadiendo un árbol al modelo aditivo, utilizando el descenso de gradiente para minimizar la función de pérdida cada vez que se añade un árbol. Para ello, se parametriza el nuevo árbol, modificando los parámetros del mismo para reducir las pérdidas residuales. Finalmente, la salida de este árbol se añade a las salidas de los otros árboles que ya formaban parte del modelo, intentando así corregir o mejorar la salida final del modelo.

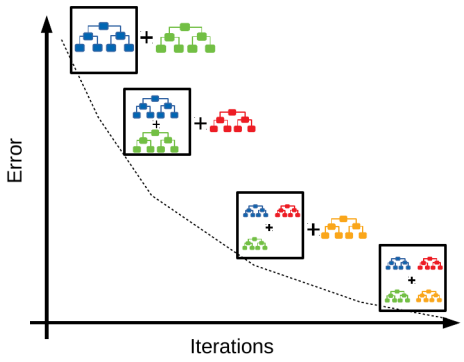

Figura 2: Representación gráfica del funcionamiento del algoritmo del Gradient Boosting.

La librería que se utilizó para los algoritmos de los árboles de decisión fue *Scikit-learn* [6]. Esta librería cuenta con una gran variedad de algoritmos y utilidades, que van desde el preprocesado de los datos hasta la evaluación de los modelos entrenados. *Scikit-learn* es una librería especializada en el Machine Learning y que tiene implementados todos los algoritmos de predicción que fueron utilizados a lo largo del trabajo, por eso fue la librería elegida para desarrollar los algoritmos que se utilizaron para el análisis.

Además, para analizar la influencia de las variables sobre el exitus de las pacientes y cómo varía el exitus según los distintos valores, se han utilizado técnicas de estadística clásica. En primer lugar, se ha empleado el Test exacto de Fisher [7]. Para aquellas variables categóricas no binomiales, se ha utilizado el test comparando la distribución de cada una de las variables contra la distribución de todas las restantes juntas. Para aplicar este test, se crea una tabla de contingencia que contiene los datos de las categorías que se quieren comparar.

	Columna 1	Columna 2	Total Fila
Fila 1	a	b	a + b
Fila 2	c	c	c + d
Total Columna	a + c	b + d	a + b + c + d (= n)

Tabla 1: Tabla de ejemplo para la aplicación del Test de Fisher.

Con una tabla como la mostrada en el ejemplo, la probabilidad de obtener ese conjunto específico de valores de una misma distribución viene dada por

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)\,!\,(c+d)\,!\,(a+c)\,!\,(b+d)\,!}{a\,!\,b\,!\,c\,!\,d\,!\,n\,!} \tag{1}$$

Con este test, será posible saber si hay algún valor de las variables que afecte más al exitus de la paciente. Para analizar si esta influencia es positiva o negativa en comparación con los otros valores de la variable, se ha utilizado el intervalo de confianza binomial según la modificación del intervalo de Clopper-Pearson [8] propuesta por Lancaster [9]. Este cálculo permite conocer el intervalo de error del exitus en ese valor de la variable. De este modo, para aquellos valores para los cuales se observe que existe una influencia en el exitus final, comparando los intervalos de confianza binomiales con los otros valores de las variables permitirá comprobar si la supervivencia de la paciente aumenta con este determinado valor o si por el contrario disminuye.

3. Curación de datos

Como en todos los proyectos de Data Science, uno de los pasos más importantes que se tienen que dar es la curación de datos. Este proceso es vital dentro del proyecto, ya que le proporciona un gran valor añadido a los datos de los que se dispone, evitando que se lleguen a conclusiones erróneas por posibles errores en los datos y mejorando la calidad de los mismos. Además, en este trabajo la curación y la limpieza de los datos cobran una mayor importancia si cabe, ya que se cuenta con muy pocos datos y todos y cada uno de ellos van a ser necesarios para el resultado final.

3.1. Limpieza de los datos

Como se mencionó en la Sección 2, la librería que se utilizó para cargar los datos del Excel a Python y tratarles es *Pandas*. *Pandas* cuenta con una gran ventaja con respecto a otras librerías y es que, por defecto, interpreta el tipo de cada columna basándose en los datos que tiene. Sin embargo, en esta ocasión no ha reconocido del todo bien las columnas que contienen fechas, como pueden ser la fecha de nacimiento de la paciente o la fecha en la que se realizó la cirugía. Esto se debe a que todas las columnas tenían formatos de fecha muy diferentes entre sí dentro de sus propios datos. Por ejemplo, había algunas fechas en el que el año venía con dos dígitos y otras en las que venía con cuatro. Por lo tanto, se tuvo que realizar una curación a todas las columnas de fechas. Para ello, se utilizó la función *to_datetime* de *Pandas*, que permite pasar uno o más datos al tipo fecha de *Pandas*. Así pues, se aplicó esta función a todas las columnas de tipo fecha, poniendo a True el parámetro *dayfirst* de la función para que interprete las fechas con el día primero.

Al acabar de pasar todas las columnas de fechas al tipo fecha de *Pandas*, se realizaron una serie de comprobaciones para confirmar que todas las fechas se habían formateado bien. El principal problema fue con las fechas que no tenían los cuatro dígitos en el año, ya que algunas se habían leído con 100 años de diferencia. Para ello, se comprobó que todas las fechas de las pacientes fueran anteriores al año 2000, ya que es el año a partir del cuál esto comienza a ser problemático. Además, comprobando la edad de las pacientes a la fecha del diagnóstico se confirmó que todas ellas tenían más de 20 años, por lo que todas habían nacido antes de este año. A todas aquellas fechas de nacimiento posteriores al año 2000 se les restaron 100 años, para que así coincidieran con las fechas de nacimiento reales. Para todas las demás fechas, se hizo una comprobación similar. Todos los datos de las pacientes son de pacientes que fueron operadas en los últimos años, por lo que se comprobó que todas las demás fechas fueran posteriores al año 2000. Al igual que con las fechas de nacimiento, a aquellas que no lo fueran se les sumó 100 años para que concordaran con la fecha verdadera.

Aparte de la curación de las fechas, se cambió el tipo de la columna y los datos en todas las columnas con variables categóricas. Estas columnas estaban categorizadas por números en el fichero Excel, así que *Pandas* las hizo de tipo numérico flotante de manera predeterminada. Como el tipo categórico de *Pandas* sólo admite valores de texto como categorías, y es interesante tener estas columnas categorizadas como números, todos los datos se pasaron a números enteros, y el tipo de las columnas

a tipo numérico entero.

Otra de las curaciones que se decidió hacer fue la normalización de la categorización de todos los datos binomiales para que las únicas categorías disponibles fueran 0 o 1, ya que algunos de ellos estaban categorizados como 1 y 2. De esta forma, el 0 significa normalmente la ausencia de esa variable, ya sea porque no se realizó ese determinado tratamiento o no se detectó esa variable, mientras que el 1 significa la presencia de la variable, implicando que el tratamiento sí se realizó o la variable sí que se detectó. Para el exitus, que es la variable objetivo de este trabajo, se determinó que el 0 significara que la paciente ha fallecido, mientras que el 1 significara que la paciente sigue viva. La curación de esta columna fue particularmente importante, no sólo por la relevancia de la variable, si no también porque, aunque estaba categorizada como 1 y 2, existían algunos valores que eran 0 para algunas pacientes. Por lo tanto, se utilizó la columna de la fecha del exitus de la paciente para determinar el valor de la variable objetivo. En el caso de que no haya fecha de exitus, significaría que la paciente sigue viva, mientras que si hay fecha de exitus, la paciente habría fallecido. Esta fue la condición que se utilizó para curar los datos de esta columna.

Por último, se realizó también una limpieza de los datos de la columna que contiene el número de ciclos aplicados en la quimioterapia. Esta columna era un poco especial, ya que a algunas pacientes se les había aplicado más de un tipo de quimioterapia, por lo que esta columna contenía el número de ciclos que se había aplicado para cada una de las quimioterapias realizadas, separados por distintos símbolos. Finalmente se optó por sumar el número de ciclos de cada una de las quimioterapias aplicadas para normalizar la columna y que sólo tuviera números enteros.

3.2. Preprocesado de los datos

Una de las partes fundamentales dentro del proceso de la curación de los datos es el tratamiento de los datos vacíos. En un proyecto de Data Science, es muy distinto que un dato esté vacío a que sea 0, y en este trabajo existen muchos datos que están vacíos, ya que no se dispone de esos datos. Si la cantidad de datos fuera mayor, una curación sencilla para los datos vacíos es simplemente eliminar todas aquellas filas que contengan datos vacíos. Con esto, se elimina este problema a la hora de realizar el análisis. Sin embargo, este es un proyecto atípico debido al número tan pequeño de datos con el que se cuenta. Por lo tanto, los datos vacíos tuvieron que ser tratados para no reducir aún más los datos de los que se dispuso.

Para empezar, había algunas de las columnas en las que la ausencia de dato sí que se podía considerar como 0. Por ejemplo, en la columna de Quimioterapia, se consideró que si el dato no estaba disponible era porque la paciente no había recibido quimioterapia. Otros ejemplos pueden ser todas las columnas que indican el tratamiento que se usó para las recaídas (locales, regionales y a distancia). Es bastante evidente en este caso que si no se indica nada es porque no hubo recaída, por lo que no se le aplicó ningún tratamiento. En estos casos, se trata de una curación sencilla. Sin embargo, el mayor problema viene con aquellas en las que el dato vacío no es 0, que son principalmente aquellas en las que no se hizo la prueba correspondiente para obtener el dato, o sí se hizo pero no se pudo sacar

nada concluyente.

En esta ocasión, se optó por usar un tratamiento de los datos vacíos llamado Target Encoding. El Target Encoding consiste en asignarle un número entero a cada una de las categorías de la variable. Es importante tener en cuenta que la asignación de estos números no se realiza de forma arbitraria, si no que se asignan los números gradualmente de menor a mayor gravedad de la categoría. De esta forma, este tratamiento es óptimo para aquellas variables categóricas que estén categorizadas gradualmente. En este caso, esto significaría que si hay cuatro categorías de una variable, la primera sería la menos grave y la cuarta la más grave, ascendiendo progresivamente (o al revés). Por ejemplo, una de las variables para las que el Target Encoding encaja perfectamente es el Estadio FIGO (Federación Internacional de Ginecología y Obstetricia) prequirúrgico de la paciente, que es el estadio del tumor visto por una resonancia magnética antes de operar. En esta ocasión, un dato vacío no se puede sustituir por un 0, ya que el 0 significa que no se vio ningún tumor cuando se hizo la resonancia, y un dato vacío es que no se conoce el estadio FIGO, ya que no se realizó la resonancia magnética por diversos motivos. Además, existen 9 categorías del estadio FIGO, siendo 0 la menos grave y 8 la más grave, por lo que es un ejemplo perfecto en el que aplicar Target Encoding.

Para tratar los datos vacíos, el Target Encoding se ha utilizado creando dos nuevas categorías, siendo una de ellas la menos grave y la otra la más grave. Así pues, todos los datos vacíos en los que la paciente haya fallecido se introducirían en la categoría más grave, mientras que todos aquellos en los que la paciente sigue viva irían en la categoría menos grave. De esta forma, se mantiene la categorización progresiva según la gravedad de la categoría, y se deja de contar con datos vacíos. En el ejemplo de la columna Estadio FIGO, se crearían dos nuevas categorías, -1, en donde irían todos los datos vacíos con exitus de la paciente 0, y 9, en donde irían todos los datos vacíos en los que la paciente falleció. Este tratamiento tiene una serie de inconvenientes y, como veremos luego, en algunas columnas se decidió cambiar a otro tratamiento. El principal problema es que, al contar con tan pocos datos, si existen muchos datos vacíos se podría falsear la importancia de la variable, ya que se crean dos nuevas categorías en las que la gravedad de la variable es extrema, mientras que puede que las otras categorías de las que sí se dispone información no aporten tanto por sí mismas al algoritmo.

Además de esto, se decidió realizar un One Hot Encoding para las columnas categorizadas en las que no se pudo aplicar Target Encoding. Éstas eran aquellas en las que las categorías no seguían una categorización progresiva, si no que eran categorías en las que no había una relación de continuidad entre ellas. Entre ellas se encuentran por ejemplo los lugares en los que hubo recaída del tumor, o los tipos de quimioterapia que recibió la paciente. Las diferentes categorías de estas variables no tienen una categorización continua, ya que cada una de ellas no tiene una relación con la anterior o la siguiente categoría. Lo que hace el One Hot Encoding es crear una columna para cada una de las categorías que existen, y rellenarla con un 0 o un 1; un 0 si esa categoría no está presente y un 1 si lo está. De esta forma, no es necesario tener datos en forma de array en las columnas, y se puede ver de forma muy clara si la paciente recibió o no el tratamiento o si tuvo o no recaída en ese lugar. Sin embargo, no es recomendable aplicar el One Hot Encoding a columnas que tengan muchas categorías, ya que al crear una columna por cada una de las categorías, aumentaría mucho

la dimensionalidad de los datos.

El último tratamiento de datos vacíos que se hizo fue para las columnas que contienen variables continuas, como puede ser el tamaño del tumor o Ki67, que es una forma de medir el crecimiento de las células cancerígenas. Que haya datos vacíos en estas columnas se debe a que no se pudo medir el tumor por cualquier razón, o que no se hizo la prueba de Ki67. En este caso, al ser variables continuas, se optó por utilizar un Mean Encoding. El Mean Encoding consiste en rellenar todos los datos vacíos con la media de los datos de la columna. Esta es una forma sencilla de rellenar los datos vacíos, ya que se considera la media como un valor neutral para el resultado, es decir, que se intenta decir al árbol que para esa paciente no se sabe si esa variable es beneficiosa o perjudicial.

No obstante, cuando se introdujeron los datos al algoritmo se confirmó lo mencionado anteriormente con el Target Encoding. Al crear dos nuevas categorías con valores extremos del exitus, en muchas ocasiones se falseaba la importancia que tenía la variable, ya que esas dos nuevas categorías aportan mucha más información de la que aportan las otras categorías por sí solas. Esto era especialmente notable en las columnas Estadio FIGO y p53 cruces, llegando la primera a alcanzar más del 50 % de la importancia normalizada en los 3 algoritmos utilizados. Por lo tanto, se decidió emplear un nuevo método de curación de los datos vacíos en estas dos columnas. La técnica que se utilizó consiste en rellenar los datos vacíos con un valor aleatorio de las categorías existentes, pero dándole a cada una de las categorías un determinado peso, basado en el número de pacientes que tiene la categoría. Esto se hizo de forma separada para los datos vacíos de las pacientes que habían sobrevivido y de las pacientes que habían fallecido. De esta forma, si hay una de las categorías en las que el número de pacientes fallecidas es mayor, será más probable que un dato vacío de una paciente fallecida pertenezca a esa categoría que a otra con un número menor de pacientes fallecidas.

3.3. Creación de nuevas columnas

Otra de las curaciones que se hicieron fue la creación de nuevas columnas que aporten una mayor información al algoritmo. Como en esta ocasión las fechas no aportan tanta información al algoritmo, se generaron dos columnas que tienen más importancia y que pueden ser interesantes para el análisis. Estas dos columnas se llamaron Tiempo de Espera y Tiempo hasta Exitus. La primera columna se obtiene a partir de la diferencia entre la fecha de la cirugía y la fecha de diagnóstico, y contiene el número de días que la paciente tuvo que esperar desde que se la realizó el diagnóstico hasta que se hizo la cirugía. Esta columna puede ser de interés para el análisis posterior, ya que podría ser que un mayor o menor tiempo de espera influyera en el exitus de la paciente. La segunda columna contiene el tiempo que pasó desde que se realizó la cirugía hasta que la paciente falleció. En aquellas pacientes no fallecidas, el dato de esta columna es un NotaTime (NaT), que es el valor que asigna *Pandas* a aquellas columnas de tipo fecha o diferencia de fecha en las que el dato está vacío. Esta columna no fue utilizada finalmente en el trabajo, pero puede ser útil si en un futuro alguien decide usar estos datos para otro estudio distinto.

Al analizar un poco los datos de la columna Tiempo de Espera, se observó que había algunas cosas

que no concordaban. En algunas pacientes, el Tiempo de Espera era negativo, lo que quería decir que la fecha de la cirugía de la paciente era anterior a la fecha de diagnóstico. A priori no estaba claro si esto tenía algún sentido médico, pero después de consultarlo, quedó claro que se trataba de un error humano a la hora de apuntar las fechas. Así pues, hubo que revisar los historiales clínicos de las pacientes para actualizar las fechas problemáticas. Esto llevó a hacer otras comprobaciones de fechas, como que el inicio del tratamiento de radioterapia o de quimioterapia fuera anterior a su fin.

Además, una vez avanzado con el análisis, se estimó que podría ser de interés añadir una nueva columna llamada RTE + QT. Esta columna contiene un valor indicando cuántos de los dos posibles tratamientos iniciales, radioterapia y quimioterapia, se han realizado. La columna sólo puede contener tres valores distintos; un 0 si no se ha realizado ni radioterapia ni quimioterapia, un 1 si se ha realizado una de las dos y un 2 si se hicieron los dos tratamientos.

4. Entrenamiento y Resultados

4.1. Análisis preliminar de las variables

Una vez finalizada la curación de los datos y antes de comenzar con los algoritmos de predicción, se realizó un breve análisis exploratorio de los datos para tratar de tener un idea inicial de qué variables pueden afectar más a la supervivencia de la paciente. Así pues, se decidió hacer un mapa de correlación para los dos grupos de variables que se van a estudiar, de tal manera que se pueda examinar qué variables pueden estar más correlacionadas con el exitus de la paciente para así saber qué esperar del análisis más a fondo posterior.

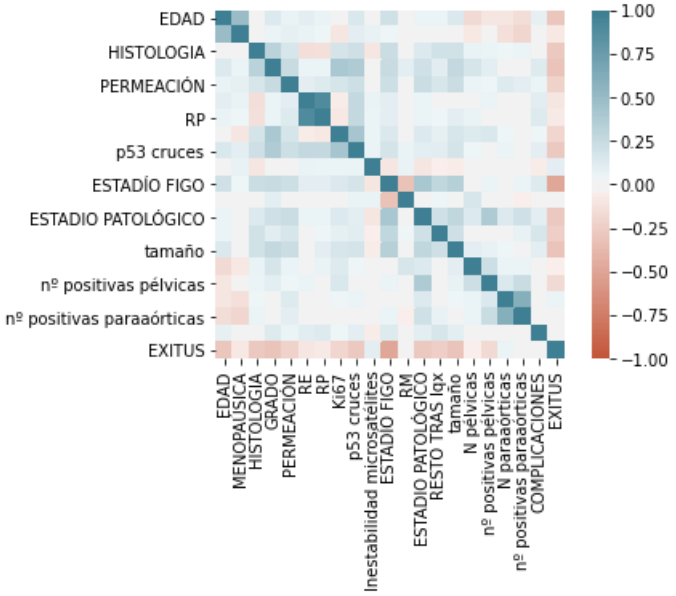

Figura 3: Mapa de correlación de las características no controlables de la paciente y el tumor, correspondientes a la primera categoría de datos.

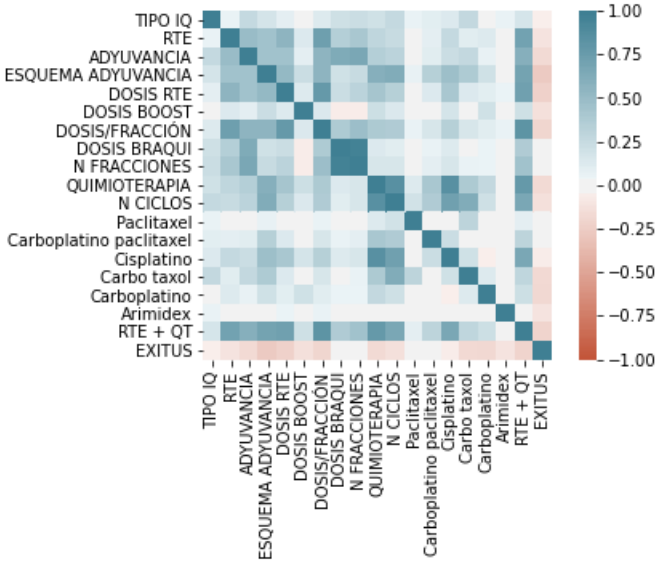

Figura 4: Mapa de correlación de las variables controlables del tratamiento, correspondientes a la segunda categoría de datos.

En la Fig. 3, se puede observar cómo hay una variable que tiene una correlación claramente más alta que las demás con el exitus de la paciente, que es Estadio FIGO. Esta es la única de las variables que destaca por encima de las demás en su correlación con el exitus, aunque tampoco tiene una correlación extremadamente alta. Además, existe alguna otra variable como puede ser la Histología, el Estadio Patológico o el Tamaño del tumor que también tienen una correlación considerable con el exitus de la paciente. En cualquier caso, con este primer vistazo, se podría estimar que seguramente la variable Estadio FIGO vaya a tener bastante peso a la hora de definir la supervivencia de la paciente.

En cambio, como se puede apreciar en la Fig. 4, el segundo grupo de variables no deja a priori ninguna variable con una correlación relevante con el exitus. La mayoría de ellas tienen una correlación baja, siendo en muchas ocasiones cercana a 0. Sin embargo, esto no quiere decir que estas variables no vayan a tener ningún peso en este estudio. Como se ha comentado en la Sección 1, el objetivo del trabajo es determinar cuáles de las variables de este segundo grupo pueden mejorar el pronóstico de la paciente en función del diagnóstico inicial. Por lo tanto, es posible que, una vez hecha una diferenciación según el diagnóstico de la paciente, estas variables cobren una mayor importancia en la decisión final.

Además de esto, se realizaron dos diagramas de barras o histogramas para cada una de las variables. Uno de ellos muestra la distribución en esa variable de las pacientes que habían sobrevivido, mientras que el otro muestra la distribución de las pacientes que habían fallecido. Las dos gráficas se muestran superpuestas, de tal forma que se pueda identificar de una manera sencilla si existe algún valor de la variable para el cual las pacientes tienen una supervivencia manifiestamente distinta de los demás valores. A continuación, se muestran algunas de las figuras más significativas que se obtuvieron con

estas visualizaciones. En todas ellas, cada uno de los diagramas de barras o de los histogramas se encuentran normalizados.

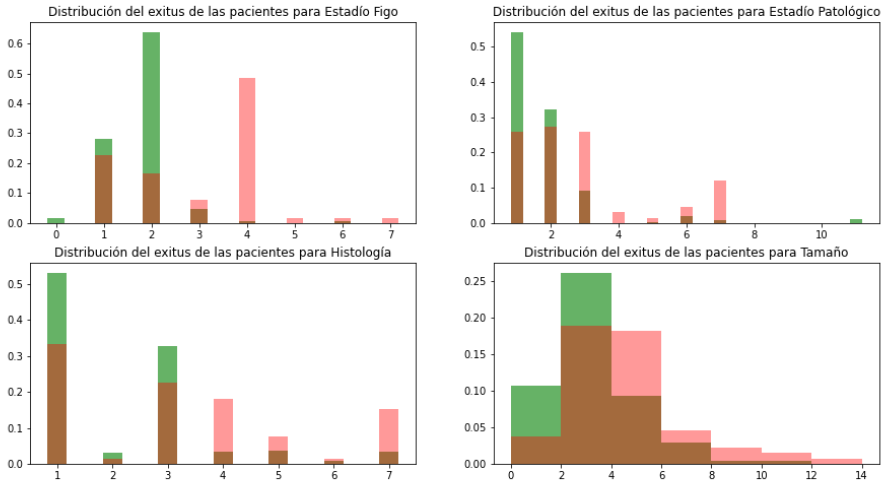

Figura 5: Histogramas y diagramas de barras normalizados de cuatro de las variables del dataset, divididas según el exitus de las pacientes.

Con un breve estudio de la Fig. 5, se puede llegar a entender el porqué de las altas correlaciones de estas variables con el exitus de la paciente. En las cuatro, aunque esto es especialmente claro en Estadio FIGO, se podría imaginar una división a partir de un determinado valor que asignara a un lado exitus 1 y a otro exitus 0, y probablemente se obtendría una buena predicción. Es importante destacar que tanto los histogramas como los diagramas de barras se encuentran normalizados, por lo que es posible que en algunos de ellos, aunque una barra se encuentre más alta que la otra, en realidad haya más datos de la segunda debido a que hay más pacientes con ese exitus. En cualquier caso, esta figura confirma la alta correlación de todas estas variables con el exitus de la paciente.

4.2. Definición de los árboles de clasificación

Una vez realizado el análisis exploratorio inicial, se comenzó con el entrenamiento de los tres métodos seleccionados para el análisis. Como se mencionó en la Sección 2, los algoritmos que se van a utilizar a lo largo del trabajo son tres. Uno de ellos es el árbol de clasificación, que crea divisiones del conjunto de datos tratando de minimizar la heterogeneidad del mismo con estas divisiones. Los otros dos están basados en este algoritmo. El primero de ellos es Random Forest, que utiliza la técnica de Bagging para mejorar la precisión final. Para ello, crea varios árboles de clasificación en paralelo, y toma la decisión final en función de la predicción que dan la mayoría de estos árboles. El otro algoritmo que se va a utilizar es el Gradient Boosting, que se basa en la técnica de Boosting.

Esta técnica también crea varios árboles de clasificación, pero en lugar de hacerlo en paralelo, va añadiéndoles para minimizar una función de pérdida.

Para empezar, se realizó una cross validación de cada uno de los métodos para tratar de encontrar los parámetros más adecuados en cada uno de ellos, especialmente el número de nodos finales. Para ello, se utilizó una cross validación de 5 folds, utilizando la media de la precisión obtenida para la gráfica de la evolución de la precisión. Se realizó una cross validación de 5 folds para un rango de nodos finales de los árboles entre 2 y 30. De este modo, se obtiene una precisión para cada uno de los valores de los nodos finales, comprobando así qué valor es más adecuado para utilizar en el algoritmo. Una cross validación de 5 folds obtiene 5 precisiones distintas del algoritmo. Para ello, divide el conjunto de datos en 5 subconjuntos de igual tamaño, y en cada una de las pruebas utiliza uno de ellos como datos para el testeo, y el conjunto restante para el entrenamiento. De esta forma, se obtiene más de un resultado para la precisión del algoritmo y este no es tan dependiente del conjunto de datos seleccionado. De todas formas, al ser el conjunto de datos con el que se está tratando tan pequeño, es posible que haya algunas pequeñas variaciones en función de los subconjuntos de datos que se seleccionen.

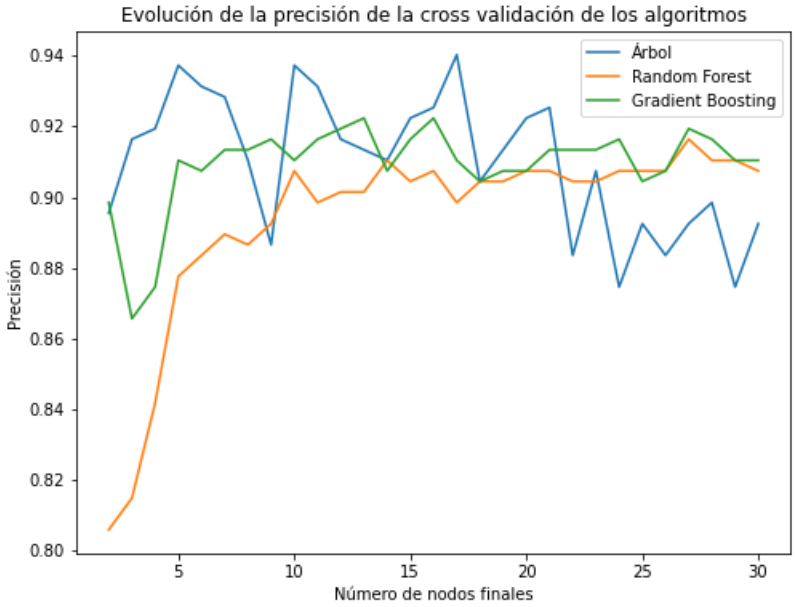

Figura 6: Evolución de la precisión de los algoritmos utilizados en función del número de nodos finales.

Como se puede observar, los tres algoritmos tienen alguna fluctuación en su precisión, ya que ni su subida ni su bajada es constante. Sin embargo, los tres tienen un pico de precisión alrededor de 10 nodos finales. Por lo tanto, el orden de magnitud del número de nodos finales es bastante claro, y para tener en cuenta las posibles variaciones debido a las divisiones del conjunto de datos, se ha seleccionado 10 como número de nodos finales para los tres algoritmos.

Así pues, una vez elegidos los parámetros de cada uno de los algoritmos, se comenzó con su entrenamiento. Para el entrenamiento, se utilizó el conjunto completo de los datos, fundamentalmente por dos razones. La primera de ellas es el tamaño del conjunto de los datos. Si se dividieran los datos en dos subconjuntos de entrenamiento y testeo, es muy probable que la división ocultara alguna información para el entrenamiento del árbol, ya que reduciría aún más el número de datos y puede que algunos datos que aporten información para el entrenamiento del algoritmo se encuentren en el conjunto de test. Además, el objetivo de este trabajo no es tanto la predicción si no el estudio de las variables que llevan a esta predicción, por lo que es de mayor interés contar con un mayor número de datos para el entrenamiento y el estudio de la importancia de las variables que utilizarles para la predicción con un modelo posiblemente incompleto debido a la falta de datos. De esta forma, sólo se realiza la cross validación para seleccionar los parámetros de cada uno de los algoritmos, pero no para su entrenamiento. El hecho de no utilizar la cross validación durante todo el proyecto es podría introducir un cierto sobre-entrenamiento en los algoritmos utilizados, pero al haber utilizado modelos sencillos para el análisis, seguramente este sobre-entrenamiento no sea excesivo.

4.3. Clasificación

Por lo tanto, utilizando todo el conjunto de datos, se entrenaron los tres distintos algoritmos presentados en la Sección 2: un árbol de clasificación, Random Forest y Gradient Boosting, todos ellos con los respectivos parámetros determinados por cross validación. La salida que dan estos tres algoritmos es un número entre el 0 y el 1, que simboliza la predicción para un determinado conjunto de las variables con las que se entrenó el árbol. Como las dos categorías para las que se predice son 0 y 1, cuanto más se acerque la predicción a uno de los dos, significa que el algoritmo está más seguro de que la predicción de ese conjunto de datos corresponde a esa categoría en concreto. Así pues, normalmente se pone como límite el 0.5 para realizar la predicción final. De esta forma, todas aquellas predicciones que tengan un valor menor a 0.5 se asignarán a la categoría 0, mientras que aquellas con un valor mayor a 0.5 se asocian a 1.

Para medir el funcionamiento de un algoritmo basado en árboles de decisión, se pueden utilizar varias técnicas. En primer lugar, es necesario definir algunos términos. El primero de ellos es el falso positivo. Un falso positivo es un dato que se predice como 1, pero su categoría real es 0. Del mismo modo, un falso negativo ocurre cuando se predice un conjunto de variables como 0, pero la categoría a la que pertenece es el 1. También se define el verdadero positivo como aquellos datos predichos como 1 cuya categoría real es el 1, mientras que un verdadero negativo es aquel conjunto de variables perteneciente al 0, y predichos como 0. Con estos cuatro términos, se definen cuatro métricas que son interesantes de analizar cuando se construyen algoritmos basados en árboles de clasificación. La sensibilidad o tasa de éxito es el número de verdaderos positivos dividido entre el número total de positivos, mientras que la tasa de falsos positivos es el número de falsos positivos dividido entre el número total de negativos. De igual manera, se define la especificidad como el número de verdaderos negativos entre el número total de negativos, mientras que el ratio de falsos negativos es el número de falsos negativos dividido entre el número total de positivos.

Así pues, con estas cuatro métricas se puede definir la curva ROC, que es una de las técnicas que se pueden usar para medir cómo de bien funciona un algoritmo de clasificación. La curva ROC se construye representando la sensibilidad del algoritmo frente a 1 menos la especificidad. Para ello, se va variando el umbral a partir del cual se considera que la predicción es positiva o negativa. De esta forma, para cada umbral que se utilice se incluye un punto en la curva ROC, ya que se van variando todas las métricas construidas. A continuación, se muestra un ejemplo de una curva ROC. Cuanto más se aleje la curve de la línea diagonal divisoria, mejor es el algoritmo, ya que la línea diagonal representaría un algoritmo totalmente aleatorio.

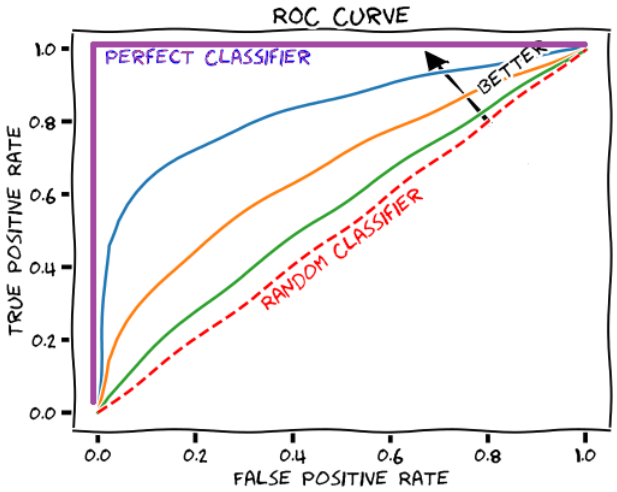

Figura 7: Ejemplos de varias curvas ROC.

Además de la curva ROC, también se puede emplear otra métrica para evaluar cómo de bien funciona un algoritmo basado en árboles de decisión. Esta métrica es la precisión, y siguiendo las definiciones de los términos expuestos anteriormente, es la suma de los verdaderos positivos y los verdaderos negativos dividido entre el número total de datos. Esta métrica da una idea más general del funcionamiento del algoritmo, ya que se fija sólo en los resultados acertados, mientras que con la curva ROC se puede entrar más en detalle en los aciertos en cada una de las categorías. A lo largo de este trabajo, la métrica que se usó fue la precisión para evaluar los distintos algoritmos.

En este primer entrenamiento, se seleccionaron solamente los dos grupos de variables que se van a estudiar en el trabajo, las correspondientes a las características del tumor y de la paciente y las controlables por el médico en el tratamiento inicial. Una de las primeras cosas que se comprobaron del entrenamiento fue la importancia de las variables en las decisiones de los tres algoritmos. Con este entrenamiento inicial, fue cuando se decidió cambiar la curación de los datos vacíos de las variables Estadio FIGO y p53 cruces, ya que como se mencionó en la Sección 3, la curación por Target Encoding falseaba la importancia de estas dos variables, especialmente en el caso de Estadio FIGO que llegaba a alcanzar un 50 % de importancia en los tres algoritmos. Una vez realizada esta nueva curación, se volvieron a entrenar los tres algoritmos. A continuación, se muestra el árbol de

clasificación obtenido con este entrenamiento en el primero de los algoritmos.

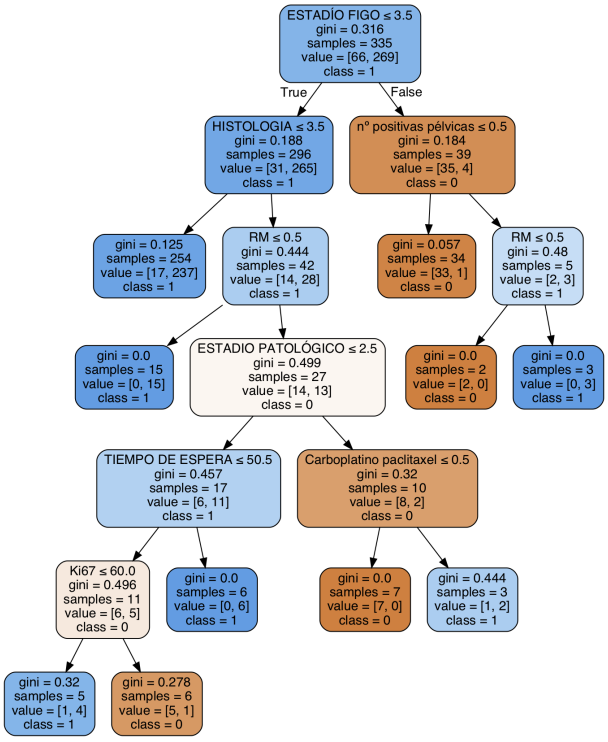

Figura 8: Representación gráfica del árbol de clasificación generado con los dos grupos de variables que se van a estudiar.

De un primer vistazo, se puede confirmar cómo el análisis exploratorio que se realizó fue bastante acertado. Tres de las cuatro variables destacadas por tener una mayor correlación con el exitus de la paciente se utilizan en el árbol para tomar una de sus decisiones. Además, las tres se encuentran bastante arriba, lo que quiere decir que tienen una mayor importancia en las divisiones del árbol que aquellas que están más abajo, ya que reducen más la heterogeneidad del conjunto de datos con esa división. Cabe destacar que el Estadio FIGO es la primera división que el árbol de clasificación realiza, lo que implica que es la variable que reduce más la heterogeneidad de la muestra en una primera instancia. También es especialmente notable cómo esta primera división parece coincidir con esa división imaginaria que se realizó para esta variable en la Fig. 5. En esta ocasión, el árbol de clasificación divide la variable Estadio FIGO a partir del valor 3.5, algo que ya se podría haber predicho a la vista de la gráfica mostrada anteriormente. Para comprobar la relevancia de las variables en los tres algoritmos, se estudió su importancia relativa dentro de los mismos. La importancia relativa de una variable da una idea de cuánto aporta esa variable a la reducción total de la heterogeneidad del conjunto de datos que realiza el árbol.

Como se puede verificar con los datos de la Tab. 2, la variable Estadio FIGO es sin lugar a dudas la

	Árbol de clasificación		Random Forest		Gradient Boosting	
1º	Estadio FIGO	0.64	Estadio FIGO	0.27	Estadio FIGO	0.45
2º	Resonancia Magnética	0.11	Edad	0.10	Tiempo de Espera	0.07
3º	Histología	0.08	Tamaño	0.07	Histología	0.06
4º	Nº Positivas Pélvicas	0.04	Estadio Patológico	0.06	Nº Positivas Pélvicas	0.05
5º	Estadio Patológico	0.04	Histología	0.06	Resonancia Magnética	0.05

Tabla 2: Cinco variables más importantes para cada uno de los algoritmos y su importancia relativa dentro de estos.

variable más importante en los tres algoritmos. Esto era algo que ya se podía esperar con el análisis exploratorio inicial que se había realizado, ya que el Estadio FIGO era con diferencia la variable que tenía una mayor correlación con el exitus de la paciente. En los tres algoritmos, esta variable se encuentra en la primera posición de la importancia relativa, teniendo un 64 % de importancia relativa en el árbol de clasificación. Además, las otras dos variables que aparecían en la representación del árbol de clasificación y se habían destacado por tener una correlación relativamente alta con el exitus de la paciente, aparecen entre las cinco variables más importantes en los tres algoritmos, exceptuando el Estadio Patológico en el Gradient Boosting. Así mismo, la otra variable que se había destacado en el análisis exploratorio inicial, el Tamaño del tumor, aparece en el tercer puesto en el algoritmo del Random Forest. También se repiten entre las cinco variables más importantes del árbol de clasificación y del Gradient Boosting dos variables como son la Resonancia Magnética y Nº Positivas Pélvicas. Ninguna de estas dos variables tenían una correlación excesivamente alta con el exitus de la paciente. Sin embargo, si se analiza con detalle la Fig. 3, se puede observar cómo estas dos variables tienen unas correlaciones bastante altas con dos de las variables ya mencionadas, y que sí tienen una elevada correlación con el exitus de la paciente. En el caso de la Resonancia Magnética, está claramente correlacionada con el Estadio FIGO, mientras que Nº Positivas Pélvicas también tiene bastante correlación con el Estadio Patológico del tumor.

Para conocer bien las variables y entender mejor las divisiones del árbol, se realizaron varias gráficas mostrando las salidas de los algoritmos en función de los valores de cada una de las variables. A estas gráficas se le añadió una línea vertical en el punto a partir del cual el árbol de clasificación realiza la división del conjunto. De este modo, se puede observar cómo se distribuyen las salidas frente a las variables en los distintos algoritmos, y comprobar el sentido de las divisiones que realiza el árbol de clasificación.

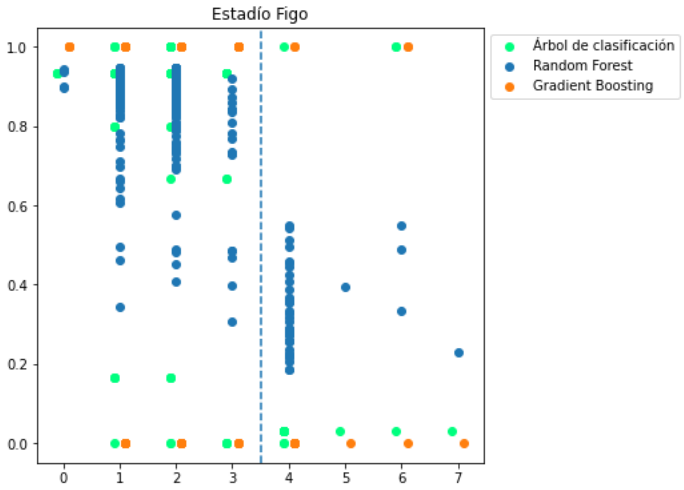

Figura 9: Distribución de las salidas de los algoritmos en función de los valores de Estadio FIGO. Los puntos del árbol de clasificación y del Gradient Boosting se han movido 0.1 hacia la izquierda y hacia la derecha respectivamente para evitar el solapamiento.

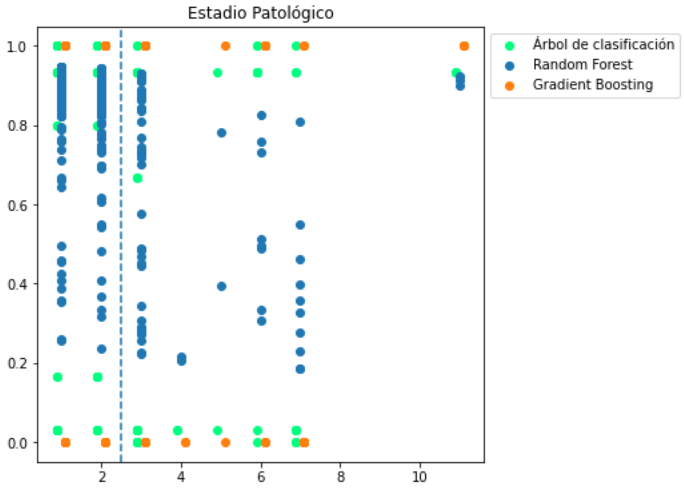

Figura 10: Distribución de las salidas de los algoritmos en función de los valores de Estadio Patológico. Los puntos del árbol de clasificación y del Gradient Boosting se han movido 0.1 hacia la izquierda y hacia la derecha respectivamente para evitar el solapamiento.

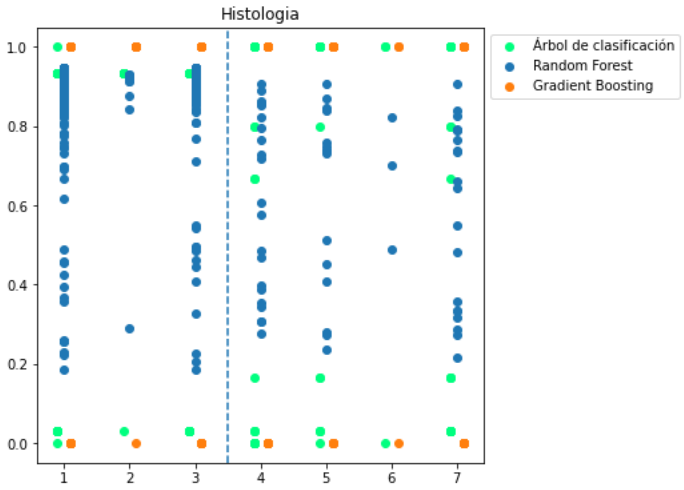

Figura 11: Distribución de las salidas de los algoritmos en función de los valores de Histología. Los puntos del árbol de clasificación y del Gradient Boosting se han movido 0.1 hacia la izquierda y hacia la derecha respectivamente para evitar el solapamiento.

Si bien para las variables de Histología y Estadio Patológico estas figuras no aportan tanta información, para el caso de Estadio FIGO queda meridianamente claro que esa división es la acertada. A la izquierda de la línea divisoria, hay una gran concentración de puntos por encima del 0.5, indicando que para los algoritmos existe una gran tasa de supervivencia a partir de esa línea. Sin embargo, a la derecha de esta división casi todos los puntos se encuentran en la parte inferior del gráfico, lo que indica que los tres algoritmos consideran que cualquier paciente con alguno de esos valores de Estadio FIGO va a tener una tasa de supervivencia mucho menor. También era de esperar que la información que aporten estas figuras para las otras dos variables no fuera tanta como para el Estadio FIGO. En esta variable, la división que se realiza es la división inicial que aplica el árbol, por lo que era esperable que con el total del conjunto de datos se viera un patrón claro. Sin embargo, las divisiones que realiza el árbol en las otras dos variables son en puntos más bajos del árbol, por lo que a esas divisiones solamente llega un subconjunto más pequeño de los datos, lo que hace que sea más difícil de visualizar la división con todos los datos.

Para terminar de analizar estas variables de diagnóstico, se analizaron sus distribuciones fijando los distintos valores de Estadio FIGO. De esta manera, se puede comprobar si las divisiones que el árbol realiza tienen sentido, y también cómo influye el valor de Estadio FIGO en los valores de las distintas variables. A continuación, se muestran las distribuciones de estas variables para dos de los valores de Estadio FIGO. Las demás gráficas se pueden encontrar en el Anexo A.2.

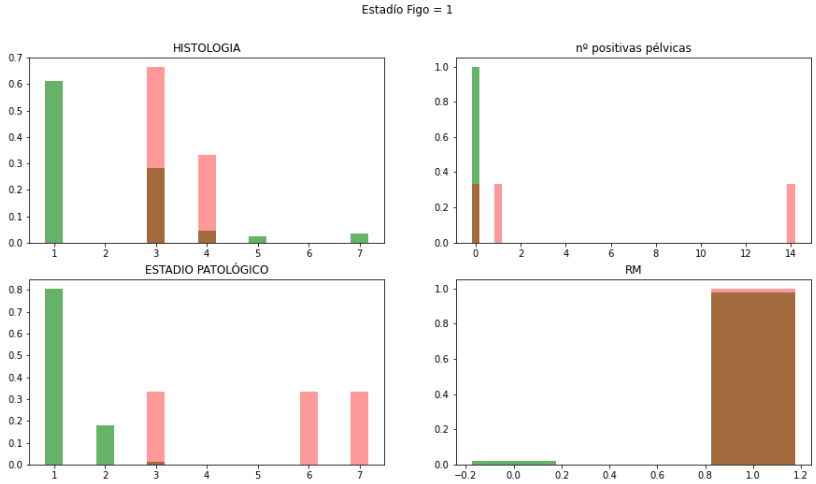

Figura 12: Distribución de las distintas variables de diagnóstico divididas según el exitus para el valor de Estadio FIGO 1. El color verde representa aquellas pacientes que sobrevivieron al tumor, mientras que el color rojo representa aquellas pacientes fallecidas.

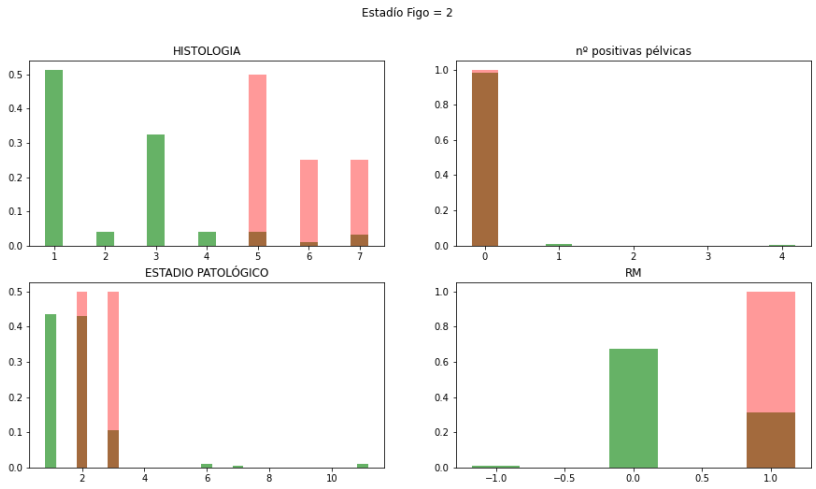

Figura 13: Distribución de las distintas variables de diagnóstico divididas según el exitus para el valor de Estadio FIGO 2. El color verde representa aquellas pacientes que sobrevivieron al tumor, mientras que el color rojo representa aquellas pacientes fallecidas.

Como se puede observar, da la impresión de que al fijar el valor de Estadio FIGO, se sigue pudiendo dividir el conjunto de datos de una forma bastante clara a partir de ciertos valores de Estadio Patológico y de Histología. Esto reafirma la idea de la importancia de estas dos variables de diagnóstico unidas con Estadio FIGO.

Además, para analizar la precisión de los algoritmos utilizados, se realizó una gráfica mostrando la distribución de la salida de los mismos en función del exitus de la paciente. Así pues, se muestran dos histogramas para cada uno de los algoritmos, en donde uno de ellos contiene la distribución de las pacientes fallecidas, mientras que el otro contiene la distribución de las pacientes que sobrevivieron al tratamiento. De esta manera, se puede reconocer de un modo muy visual si hay algún algoritmo con mayor precisión que otro, y si la predicción es mejor para las pacientes vivas o para las fallecidas.

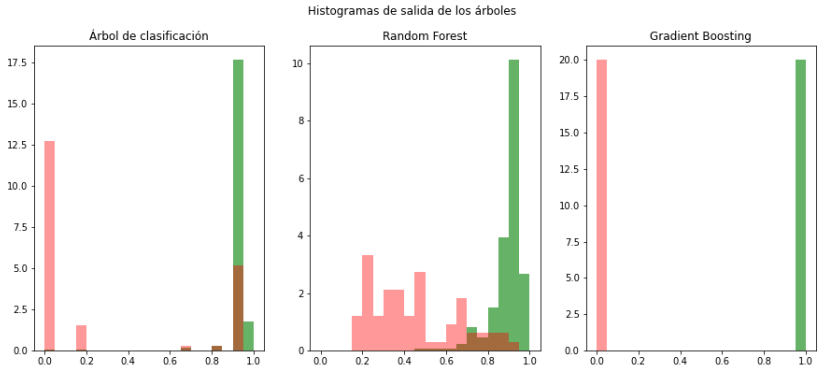

Figura 14: Distribución de las salidas de los tres algoritmos en función del exitus de la paciente.

Como se puede observar, los dos primeros algoritmos tienen una distribución más o menos similar. En ambos casos, todas las pacientes para las que el algoritmo da un valor del exitus igual a 0, son predichas correctamente, mientras que algunas en las que la predicción del exitus es 1, el exitus real de la paciente es 0. En cualquier caso, la precisión de ambos algoritmos es bastante buena, siendo un 0.93 para el árbol de clasificación y un 0.94 para el Random Forest. Para el Gradient Boosting, la situación es un poco especial. En la Fig. 14, se puede ver cómo el Gradient Boosting predice con una total exactitud todos los valores del exitus del conjunto de datos, siendo su precisión final de 1. Por lo tanto, nos encontramos ante un claro caso de over-fitting, ya que este algoritmo ha diseñado un modelo excesivamente a la medida de este conjunto de datos. Aunque el objetivo del trabajo sea el de tratar de entender cómo afectan las variables al exitus final de la paciente y no la predicción del exitus, tampoco es interesante utilizar un algoritmo que sea tan a medida que sólo sirva para estos datos con los que se ha entrenado, ya que probablemente la introducción de un solo nuevo dato cambiaría totalmente esto. De esta forma, a partir de este momento se decidió emplear solamente los dos primeros algoritmos, utilizando el árbol de clasificación más para la visualización de los cortes que realiza a las variables y el Random Forest más para el análisis de las distribuciones de las variables con el exitus.

5. Estudio de la influencia de los distintos grupos de variables

5.1. Árbol con variables de diagnóstico

Así pues, el siguiente paso que se realizó fue entrenar los dos algoritmos utilizando sólo las variables de diagnóstico de la paciente. Como se ha visto en los análisis previos de las importancias de las variables en las decisiones de los algoritmos, todas las que aparecían entre las cinco más importantes pertenecían a este primer grupo. Entrenando los algoritmos con tan sólo las variables no controlables, dará una idea de la distribución de las mismas en función del exitus de la paciente, por lo que se podrá comprobar si el uso de las variables de tratamiento puede influir en el exitus final, o si este viene determinado solamente por las variables de diagnóstico y el tratamiento que se le aplique a la paciente no tiene ninguna influencia. Para el entrenamiento, las variables de diagnóstico se han resumido en tres: Estadio FIGO, Histología y Estadio Patológico. Estas tres de las variables más importantes en los tres algoritmos, y dos de las otras variables que aparecían entre las variables con más importancia estaban bastante correlacionadas con ellas. Además, como se pudo observar en las Fig. 12 y Fig. 13, estas dos variables se complementan muy bien con Estadio FIGO para realizar divisiones claras del conjunto de datos. De esta forma, se simplifica la creación del árbol y la determinación de posibles grupos de pacientes en función de tan solo estas tres variables. El árbol de clasificación que se obtuvo con estas tres variables fue el siguiente.

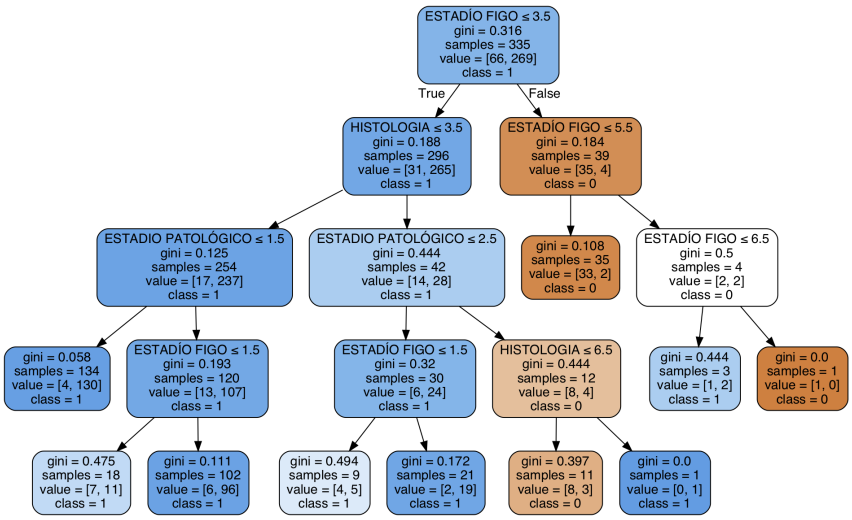

Figura 15: Representación gráfica del árbol de clasificación para las tres variables más importantes del grupo de las variables de diagnóstico.

Una vez más, la primera división que realiza el árbol de clasificación es Estadio FIGO, siendo la misma que en el árbol que contaba con todas las variables de los dos grupos de estudio. Además, hay varias divisiones más de Estadio FIGO a lo largo del árbol, lo que indica que volverá a ser la

	Árbol de clasificación		Random Forest	
1º	Estadio FIGO	0.83	Estadio FIGO	0.58
2º	Histología	0.10	Estadio Patológico	0.22
3º	Estadio Patológico	0.07	Histología	0.20

Tabla 3: Importancia relativa de las tres variables utilizadas para construir el árbol de las variables de diagnóstico.

variable más importante, algo que se confirma en la Tab. 3. En el árbol de clasificación, Estadio FIGO representa el 83 % de la importancia relativa de las tres variables, mientras que para el Random Forest representa el 58 %. A la vista de esto y de las divisiones y las hojas finales de la Fig. 15, se puede realizar una división bastante clara en dos grupos de pacientes simplemente con una división en Estadio FIGO. Para aquellas pacientes con un Estadio FIGO menor que 3.5, es decir, sin tumor o con Estadio Figo IA, IB o II, se puede observar como el árbol da un diagnóstico positivo para la mayoría de ellas, mientras que para aquellas en las que el Estadio FIGO es mayor que este valor, el diagnóstico que proporciona el árbol es negativo. Por lo tanto, se puede afirmar que existe un comportamiento distinto en las pacientes en función de las variables de diagnóstico, que se puede simplificar de una forma bastante precisa con una división en la variable Estadio FIGO. A partir de esto, el análisis se centrará en estudiar la influencia de las variables de tratamiento en el comportamiento de estos dos grupos, centrándose especialmente en el grupo con el diagnóstico inicialmente negativo y tratando de resolver si alguna de las variables de tratamiento puede influir positivamente en el exitus final de la paciente.

5.2. Estudio de las variables de tratamiento

Antes de continuar con el análisis, se decidió examinar más en detalle las variables de tratamiento que se van a estudiar en la siguiente sección. En un primer lugar, se ha comparado la influencia de los valores de Estadio FIGO en las distintas categorías de las variables de tratamiento, ya que se ha considerado la variable más importante. Para ello, se ha fijado el valor de Estadio FIGO para cada una de sus categorías, y se han representado las distribuciones de los exitus para cada una de las variables de tratamiento que se van a analizar. Para mostrar la diferencia entre dos valores de Estadio FIGO, se muestran las gráficas de las variables para los valores 1 y 4. Se han elegido estos dos valores porque son bastante indicativos de la diferencia en el exitus final de las pacientes en función de esta variable, y se puede observar bastante bien como las distribuciones de las variables de tratamiento son bastante diferentes. El resto de las gráficas se pueden encontrar en el Anexo A.3.

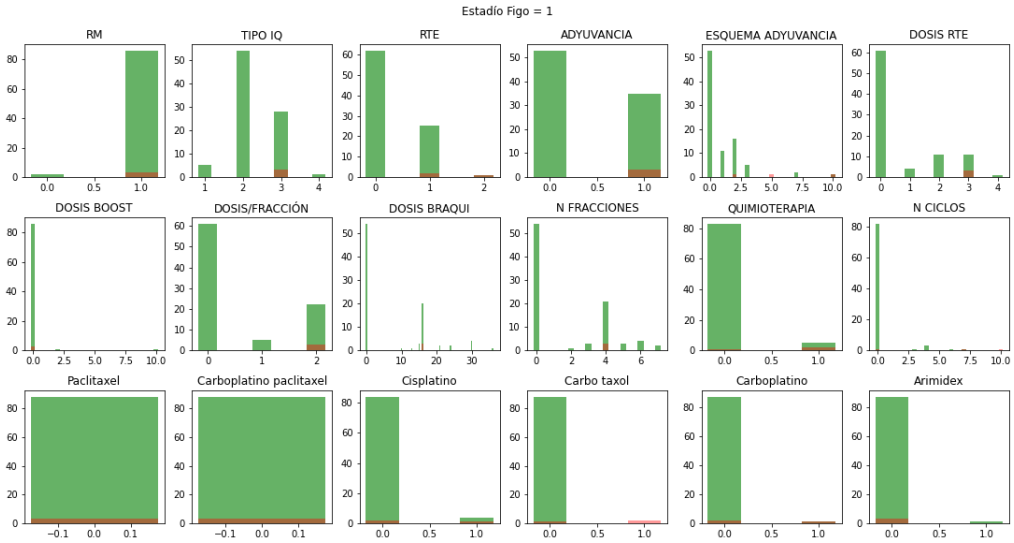

Figura 16: Distribución de las distintas variables de tratamiento separadas según el exitus para el valor de Estadio FIGO 1.

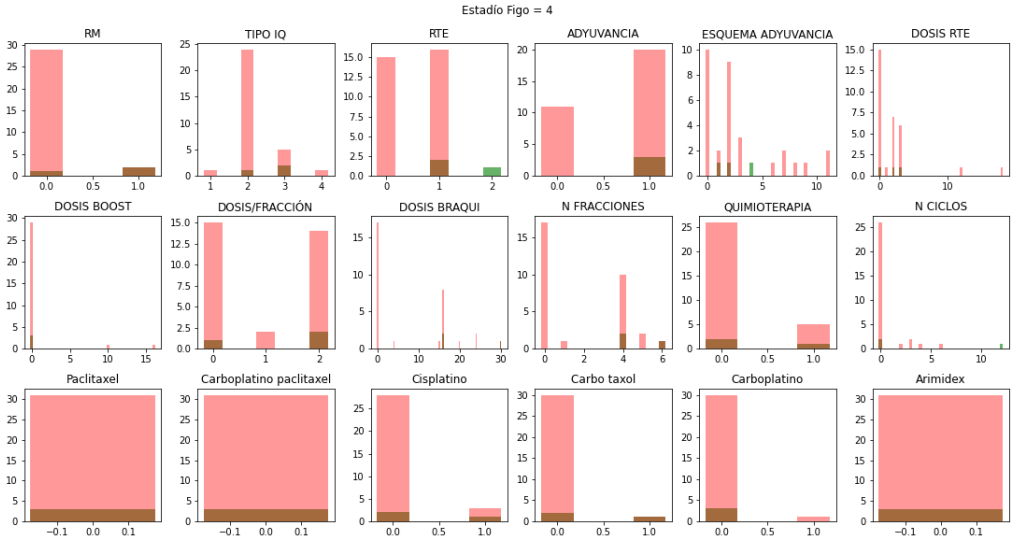

Figura 17: Distribución de las distintas variables de tratamiento separadas según el exitus para el valor de Estadio FIGO 4.

Como se puede observar, las distribuciones son realmente distintas. Una primera diferencia que se puede observar es en el número de pacientes fallecidas. Para el valor de Estadio FIGO 1, es bastante claro que existe un número mucho más bajo de pacientes fallecidas que para el valor de Estadio FIGO 4, en el que existe un número claramente más alto de pacientes fallecidas. Aparte de esto, también es bastante notable la diferencia que existe en las distribuciones de algunas de las variables, que cambian de una forma muy evidente de un valor de Estadio FIGO al otro.

Por lo tanto, se va a optar por estudiar las variables de tratamiento divididas en dos bloques. Para ello, se han representado las distribuciones de las variables de tratamiento, pero a su vez divididas en función de la salida del árbol entrenado sólo con las tres variables en las que se simplificaron las variables de diagnóstico. Es decir, todas aquellas pacientes predichas como fallecidas por este árbol se encuentran a un lado, mostrando la distribución según su exitus real, mientras que al otro lado se muestran aquellas que el árbol predijo como exitus 1. Es importante destacar que, como se verá posteriormente, si se realiza la división a partir de Estadio FIGO > 3.5, el resultado es prácticamente el mismo. A continuación, se muestran algunas de las gráficas de las variables de tratamiento. Las gráficas restantes se encuentran en el Anexo A.4.

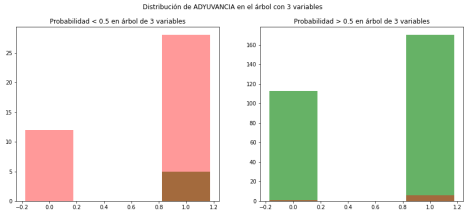

Figura 18: Distribución de Adyuvancia dividida según la predicción del árbol de las variables de diagnóstico.

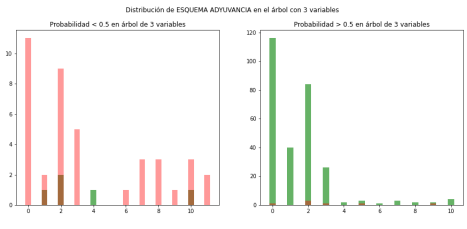

Figura 19: Distribución de Esquema Adyuvancia dividida según la predicción del árbol de las variables de diagnóstico.

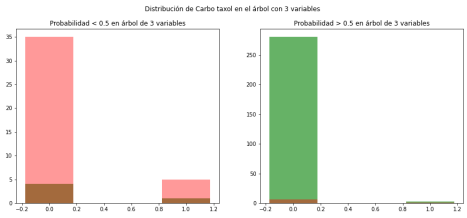

Figura 20: Distribución de Carbo Taxol dividida según la predicción del árbol de las variables de diagnóstico.

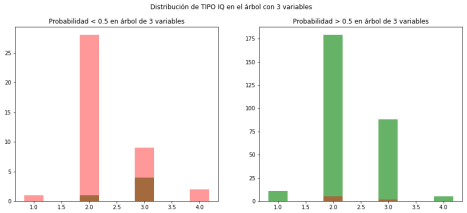

Figura 21: Distribución de Tipo IQ dividida según la predicción del árbol de las variables de diagnóstico.

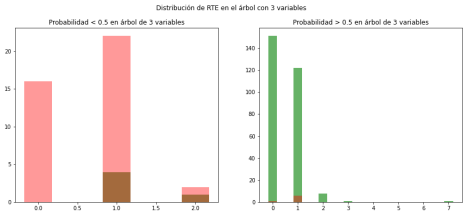

Figura 22: Distribución de RTE dividida según la predicción del árbol de las variables de diagnóstico.

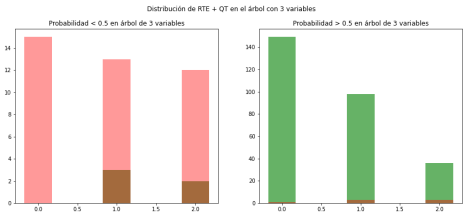

Figura 23: Distribución de RTE + QT dividida según la predicción del árbol de las variables de diagnóstico.

Según las gráficas anteriores, se puede confirmar la diferencia en la tendencia entre los dos grupos divididos según las variables de diagnóstico, por lo que parece razonable utilizar este enfoque. Además, también parece como sí que hay algunas variables de tratamiento que muestran alguna diferencia en el exitus final según sus valores, de modo que tiene sentido continuar con un análisis más a fondo de estas variables. De la misma forma, se sigue observando la tendencia para el grupo de pacientes con el diagnóstico positivo, según la que parece que la paciente tiene más probabilidades de sobrevivir si no se aplica ningún tratamiento. A partir de este punto, se planteó utilizar también técnicas de Machine Learning para continuar con el análisis. Sin embargo, si la cantidad de datos ya es reducida de por sí, al dividirlos en dos grupos se reducen aún más, especialmente para el grupo de pacientes con el diagnóstico negativo. Así pues, se decidió emplear técnicas de estadísticas convencionales para el análisis en profundidad de estas variables.

6. Análisis del efecto de los distintos tratamientos

6.1. División de las pacientes según su diagnóstico inicial

Como se ha demostrado a lo largo del trabajo, el diagnóstico inicial de la paciente tiene una gran influencia en el exitus final de la paciente. Basándose en el conjunto de las variables de diagnóstico, se pueden reducir las pacientes a dos grupos, uno para el que el diagnóstico es positivo y otro para el que el diagnóstico es negativo. Para simplificar esta división, se ha decidido reemplazar el árbol con las variables de diagnóstico por un corte en Estadio FIGO, que sería como un árbol de un solo nivel. Los resultados que se obtienen a partir de este corte son prácticamente los mismos, y se simplifica mucho esta división, siendo más fácil de implementar si alguien quisiera replicarlo.

De nuevo, antes de continuar con el análisis, se van a representar las distribuciones de las variables de tratamiento, pero divididas esta vez en función de su valor de Estadio FIGO, dividiéndolas por el valor límite identificado por las distintas gráficas y árboles. Las figuras restantes se pueden encontrar en el Anexo A.5.

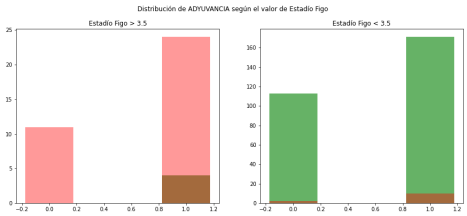

Figura 24: Distribución de Adyuvancia dividida según el valor de Estadio FIGO.

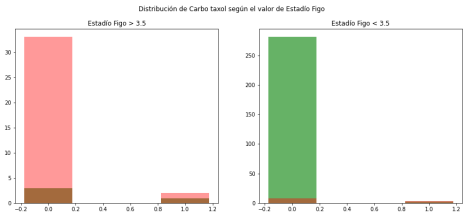

Figura 26: Distribución de Carbo Taxol divididas según el valor de Estadio FIGO.

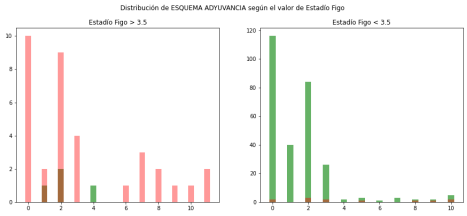

Figura 25: Distribución de Esquema Adyuvancia dividida según el valor de Estadio FIGO.

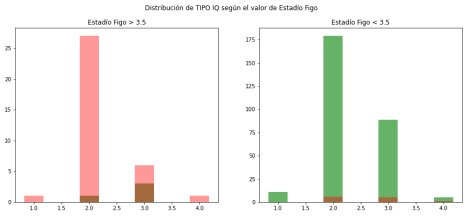

Figura 27: Distribución de Tipo IQ dividida según el valor de Estadio FIGO.

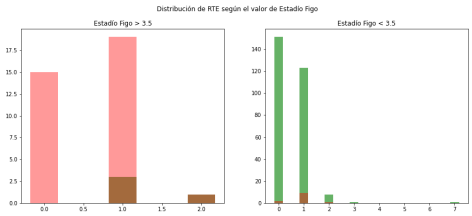

Figura 28: Distribución de RTE dividida según el valor de Estadio FIGO.

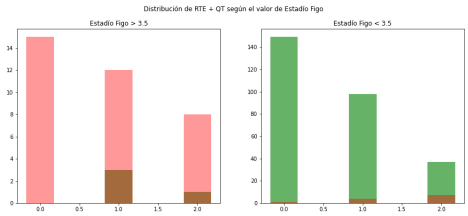

Figura 29: Distribución de RTE + QT dividida según el valor de Estadio FIGO.

Como era de esperar, la distribución de las variables según esta división a partir del valor de Estadio FIGO no varía apenas de las distribuciones divididas según la salida del árbol de las variables de diagnóstico. Además, se puede volver a observar la tendencia en el grupo de pacientes en el que el diagnóstico es positivo para la cual parece ser mejor no aplicar ningún tratamiento. Así pues, el siguiente paso es estudiar cuáles de los valores de estas variables de tratamiento tienen una mayor influencia en el diagnóstico final, y estudiar si esta influencia es positiva, aumentando la supervivencia de la paciente, o si por el contrario es una influencia negativa.

6.2. Influencia del tratamiento en las pacientes con diagnóstico inicial negativo

Para estudiar cuáles de los valores de las variables de tratamiento tienen una mayor influencia en el exitus de la paciente, se utiliza el Test de Fisher. El Test de Fisher se utiliza para variables binomiales, y obtiene la probabilidad de que las dos distribuciones vengan de la misma distribución. Para estudiar las más importantes, se van a analizar únicamente aquellas que tengan un valor inferior a 0.32, que corresponde a una sigma en una distribución gaussiana. Para p-valores menores que 0.32, se rechaza la hipótesis nula, que es que los dos valores vienen de la misma distribución, lo que implica que cambiar el valor de esa variable no tiene ninguna influencia en el exitus de la paciente. Típicamente se elige un valor límite de 0.05, correspondiente a dos sigmas, ya que proporciona una mayor seguridad en que las distribuciones no provengan de la misma distribución, pero debido a la reducida cantidad de datos con las que se cuenta, reduciéndose aún más para aquellas pacientes cuyo diagnóstico es negativo, se ha decidido emplear una sola sigma. Por lo tanto, a partir de ahora los resultados que se obtengan se considerarán más bien indicios, ya que una significación a una sigma puede significar que simplemente se trate de fluctuaciones estadísticas. Para aquellas variables multinomiales, se ha usado también el Test de Fisher para obtener el p-valor, pero se han ido comparando uno por uno cada uno de los valores de esa variable contra el resto de valores juntos. De esta forma, se comprueba si ese valor puede provenir de la misma distribución que el resto de valores.

Además, para ver si esta influencia es positiva o negativa, se calculó el intervalo de confianza binomial sólo para aquellas con p-valor menor de 0.32, según la modificación del intervalo de Clopper-Pearson [8] propuesta por Lancaster [9], el llamado mid-P interval. Este intervalo proporciona el rango, al

68.3 % en este caso, de la probabilidad de supervivencia según el valor de cada variable. Analizando este intervalo en aquellas variables con un p-valor que se haya considerado significativo, se puede resolver cuál de los valores aumentan la supervivencia de la paciente. En primer lugar, se van a estudiar las variables de tratamiento para el grupo de las pacientes para el cual el diagnóstico a partir del valor de Estadio FIGO es negativo, es decir, aquellas para las cuales el Estadio FIGO es superior a 3.5. En aquellas variables categóricas no binomiales, se va a estudiar la distribución de cada una de las categorías frente a la distribución del resto.

Variable	Valores	P-valor	Probabilidad de supervivencia (68.3 % CL)	
Adyuvancia	0	0.31	0	0.10
	1		0.09	0.22
Carbo Taxol	0	0.28	0.05	0.14
	1		0.11	0.65
Tipo IQ	2	0.06	0.01	0.09
	Resto		0.16	0.43
Tipo IQ	3	0.03	0.19	0.51
	Resto		0.01	0.09
RTE	0	0.15	0	0.07
	Resto		0.1	0.26
RTE	2	0.20	0.16	0.84
	Resto		0.05	0.14
Esquema Adyuvancia	1	0.28	0.11	0.65
	Resto		0.05	0.14
Esquema Adyuvancia	4	0.10	0.32	1
	Resto		0.04	0.14
RTE + QT	0	0.15	0	0.07
	Resto		0.10	0.26
RTE + QT	1	0.28	0.11	0.33
	Resto		0.01	0.11

Tabla 4: Probabilidad de supervivencia según el tratamiento seguido. Se presentan los intervalos al 68.3 % CL calculados según los intervalos Lancaster midP para aquellos valores de las variables de tratamiento con p-valores relevantes en el grupo de pacientes con diagnóstico inicial negativo.

Como se puede observar, en seis de las variables de tratamiento hay una diferencia significativa entre algunos de los valores que tienen. Las dos variables binomiales que se han seleccionado para estudiar en este grupo de pacientes, tienen unos p-valores que están bastante cerca del límite que se puso de una sigma. En ambos casos, los intervalos Lancaster midP se solapan un poco, aunque sí que hay uno de los valores para el cual la supervivencia aumenta. Tanto para Adyuvancia y Carbo Taxol, el valor que aumenta la supervivencia es el 1. Esto quiere decir que, para este grupo de pacientes, aplicar Adyuvancia y Carbo Taxol aumentan las probabilidades de supervivencia de la paciente. Además, esto coincide con lo que se puede observar para las Fig. 24 y Fig. 26. En ambas gráficas, se

puede observar como la aplicación de estos dos tratamientos sí que aumentan las probabilidades de supervivencia de las pacientes. Además, cabe destacar el rango del intervalo para cuando sí se aplica Carbo Taxol. Este rango es tan amplio debido a los pocos datos con los que se cuenta de pacientes con Estadio FIGO superior a 3.5 a las que se les haya aplicado Carbo Taxol.

En cuanto a las otras cuatro variables, es importante destacar una por encima del resto, que es Tipo IQ. Esta variable representa el tipo de cirugía que recibió la paciente. Para esta variable, hay dos valores que tienen un p-valor realmente significativo frente al resto de la distribución, que son el 2 y el 3. Para el caso del 2, la influencia que tiene en la supervivencia de la paciente es negativa, ya que el intervalo Lancaster midP se encuentra entre 0.01 y 0.09, mientras que para el resto de valores el intervalo está entre 0.16 y 0.43. Esto coincide con lo que ocurre para el valor 3, ya que el intervalo Lancaster midP es en este caso de 0.19 a 0.51, mientras que para el resto de la distribución es de 0.01 a 0.09. Esto se puede confirmar también con la Fig. 21, ya que cuando el valor es 3, hay un porcentaje mucho más alto de pacientes que sobreviven al tratamiento, mientras que cuando se aplica un Tipo IQ 2 el porcentaje de pacientes que fallecen es realmente alto. Por lo tanto, los datos de los que se dispone parecen indicar que para aumentar la supervivencia de una paciente que entre en este grupo, es mejor no aplicar el Tipo IQ 2 y sí aplicar el Tipo IQ 3.

Algo parecido ocurre con las variables RTE y RTE + QT. Para el caso de RTE, que representa el tipo de radioterapia que recibió la paciente (o si no lo hizo), se puede comprobar como para el valor 0, que significa no haber recibido radioterapia, la supervivencia de las pacientes se ve afectada negativamente. En cambio, la supervivencia aumenta cuando se le aplica a una paciente una radioterapia tipo Pelvis + Paraaórtica, que corresponde al valor 2. En este caso, la supervivencia de la paciente está entre 0.16 y 0.84 y para el resto de valores entre 0.05 y 0.14, mientras que si no se le aplica radioterapia a la paciente la supervivencia baja hasta estar entre 0 y 0.07. Para RTE + QT, también se puede comprobar como no aplicar ni radioterapia ni quimioterapia, que corresponde con el valor 0 de esta variable, tiene un supervivencia de entre 0 y 0.07, mientras que si se aplica una de las dos (RTE + QT = 1), la supervivencia sube hasta encontrarse entre 0.11 y 0.33. De nuevo, estas dos cosas se podían intuir viendo las Fig. 28 y Fig. 29, ya que en ambos casos los resultados obtenidos con el Test de Fisher y el intervalo Lancaster midP se ven bien representados en las gráficas mostradas.

Por último, en el caso de la variable Esquema Adyuvancia hay dos valores que afectan positivamente a la supervivencia de la paciente frente al resto. Estos dos valores son 1, que corresponde a realizar una braquiterapia, y 4, que corresponde a aplicarle a la paciente una braquiterapia y una quimioterapia concomitante. Para el caso de la braquiterapia, el intervalo de supervivencia está entre 0.11 y 0.65, y para el resto es de 0.05 a 0.14, mientras que para la braquiterapia y la quimioterapia concomitante, el intervalo de supervivencia es de 0.32 a 1, siendo de 0.04 a 0.14 para el resto de la distribución. Sin embargo, hay que recalcar que para la braquiterapia, los intervalos de supervivencia se solapan ligeramente, además de tener un p-valor bastante cercano al valor límite elegido. Por lo tanto, aunque los dos aumentan la supervivencia de la paciente, si tuviera que elegir uno de los dos esquemas elegiría la braquiterapia y la quimioterapia concomitante, ya que tiene un intervalo de supervivencia que empieza en un valor más alto, además de contener el 1 que implicaría una supervivencia total,

y tiene un p-valor más significativo que la braquiterapia. Una vez más, estos resultados podían presentirse a partir de la Fig. 25, especialmente para el valor 4, ya que, aunque son muy pocas, todas las pacientes a las que se le aplicó este esquema acabaron sobreviviendo al tratamiento.

6.3. Influencia del tratamiento en las pacientes con diagnóstico inicial positivo

A continuación, se va a realizar el mismo estudio para aquellas pacientes pertenecientes al otro grupo creado a partir de la división en Estadio FIGO. En esta ocasión, al contar con más datos de pacientes, se ha considerado un p-valor significativo todos aquellos inferiores a 0.05. De esta manera, para estudiar la influencia de las variables de tratamiento en este grupo de pacientes se consigue una significación a dos sigmas. A continuación, se muestra la tabla de todos aquellos valores de las variables de tratamiento con p-valores significativos para este grupo de pacientes, incluyendo también el intervalo Lancaster midP.

Variable	Valores	P-valor	Probabilidad de supervivencia (68.3 % CL)	
Quimioterapia	0	0.00007	0.97	0.99
	1		0.77	0.88
Cisplatino	0	0.04	0.96	0.98
	1		0.82	0.93
Carbo Taxol	0	0.00005	0.96	0.98
	1		0.25	0.62
Carboplatino	0	0.009	0.95	0.97
	1		0.25	0.75
RTE	0	0.02	0.97	0.99
	Resto		0.91	0.95
RTE	1	0.04	0.91	0.95
	Resto		0.97	0.99
Esquema Adyuvancia	10	0.03	0.51	0.86
	Resto		0.95	0.97
RTE + QT	0	0.003	0.98	1
	Resto		0.90	0.94
RTE + QT	2	0.0005	0.78	0.89
	Resto		0.97	0.99

Tabla 5: Probabilidad de supervivencia según el tratamiento seguido. Se presentan los intervalos al 68.3 % CL calculados según los intervalos Lancaster midP para aquellos valores de las variables de tratamiento con p-valores relevantes en el grupo de pacientes con diagnóstico inicial positivo.

Como se puede observar, en la mayoría de los valores de las variables de tratamiento que tienen un p-valor significativo para este grupo de pacientes, aparece el valor 0, que significa que ese determinado tratamiento no se ha aplicado. Además, en todos los casos en los que aparece, la supervivencia de

las pacientes aumenta, siendo especialmente claro en las variables Quimioterapia y Carbo Taxol. Además, tanto en la variable RTE como en RTE + QT, aparecen dos resultados para los cuales la supervivencia disminuye. Esto reafirma la tendencia comentada, ya que en ambos casos no aplicar ese determinado tratamiento incluye no hacer nada, lo que aumenta las probabilidades de supervivencia. Los resultados de la Tab. 5 confirman la tendencia observada en las figuras mostradas previamente, y es que para el grupo de pacientes en las que el Estadio FIGO es menor de 3.5, el resultado mejora si no se le aplica ningún tratamiento, lo que parece algo totalmente contra intuitivo. Sin embargo, una posible explicación de este fenómeno puede ser debido a la propia naturaleza de los datos. En realidad, los datos de los que se dispone están ya filtrados de alguna manera, ya que los médicos han tenido que ir tomando decisiones para tratar a las pacientes. De esta forma, es bastante probable que aquellos tumores que no se trataron fuera porque no tenían ningún peligro para la paciente y era realmente seguro no tratarles, mientras que aquellos a los que sí se les aplicó algún tratamiento, era porque realmente lo necesitaban, ya que el tumor era más peligroso y requería de algún tipo de tratamiento para intentar evitar el fallecimiento de la paciente. Por lo tanto, esto quiere decir que en los datos de los que se dispone, ya habría un filtro por el cual a los tumores que no son tan agresivos no se les aplicaría ningún tratamiento, mientras que aquellos que tienen un diagnóstico peor sí que recibirían tratamiento, lo que podría explicar que la supervivencia sea mayor en aquellas pacientes que no han recibido ningún tratamiento.

7. Conclusiones

En este trabajo, se ha tratado de interpretar un problema médico desde un enfoque puramente basado en Data Science, con datos obtenidos a partir de unas 340 pacientes operadas de cáncer de endometrio. Para ello, se han estudiado las mejores técnicas para abordar este problema, eligiendo finalmente un modelo mixto combinando técnicas de Machine Learning basadas en árboles de clasificación y técnicas de estadística clásica, y descartando un modelo únicamente de Machine Learning basado en redes neuronales. Este modelo se descartó debido a la reducida cantidad de datos de los que se disponía, y se decidió emplear un modelo basado en árboles de clasificación, ya que trabaja mejor con un menor número de datos y aplicaba mejor a los datos disponibles, ya que se contaba con un gran número de variables categóricas que las redes neuronales no tratan del todo bien.

Respecto a las técnicas de Machine Learning empleadas, se han implementado tres distintas soluciones basadas en árboles de clasificación sobre el lenguaje de programación Python. La primera de estas soluciones fue un único árbol de clasificación, que proporciona de forma sencilla y muy visual las divisiones que realiza sobre el conjunto de datos para llegar a una diferenciación final en dos grupos. Las otras dos soluciones combinan varios árboles de clasificación. La primera de estas dos es el Random Forest, y está basada en la técnica de bagging, que crea un número de árboles de clasificación para distintas sub-muestras, y asigna la predicción en función de la predicción de todos los árboles creados. La última solución implementada fue el Gradient Boosting, que se basa en la técnica de boosting, en donde combina distintos árboles de clasificación que van minimizando una función de pérdida.

Para aplicar estas tres soluciones, se realizó una curación de datos al dataset. Principalmente, se aplicaron distintas técnicas para la curación de columnas categorizadas, que eran las más abundantes dentro del conjunto de datos. En particular, se estudiaron técnicas como el One-Hot Encoding para aquellas columnas que no tenían una categorización progresiva, creando una columna para cada una de las categorías disponibles que indicaba la existencia o la ausencia de esa categoría, o Target Encoding, que se aplicaba a aquellas columnas con una categorización progresiva y que asignaba un número entero a cada una de las categorías, ordenándolas por su gravedad. Además, se propuso una modificación a esta técnica para curar los datos vacíos, asignándoles un valor aleatorio de las categorías ya existentes, pero teniendo en cuenta el peso de cada una de las categorías en función del número de pacientes con el mismo exitus que el dato vacío.

En un análisis exploratorio inicial de las variables, se descubrió que existían varias variables que mostraban una correlación relativamente alta con el exitus de la paciente. Utilizando los árboles de decisión, se comprobó que las principales variables según su importancia en los algoritmos eran Estadio FIGO, que describe el estadio del tumor según lo visto en una resonancia magnética previa a la operación, Histología, que corresponde con la histología o el tipo del tumor, Estadio Patológico, que sería el estadio del tumor según lo observado durante la operación, Resonancia Magnética, que indica si se realizó o no resonancia magnética a la paciente, y Nº Positivas Pélvicas, que indica

cuántos de los ganglios retirados de la pelvis son positivos. La importancia de las variables en los algoritmos se podría explicar a partir de su correlación con el exitus, ya que o bien están altamente correlacionadas con el exitus de la paciente, o lo están con variables que están correlacionadas con el exitus.

Para intentar discriminar los mejores tratamientos que aplicar a las pacientes, se intentó dividir las variables en dos grupos, llamados de diagnóstico y de tratamiento, de tal forma que se separara la dependencia de este último grupo de variables de las variables de diagnóstico que parecen tener un mayor peso. Se comprobó qué utilizando un árbol de clasificación basado en las tres variables de diagnóstico con mayor importancia, Estadio FIGO, Histología y Estadio Patológico, separaba dos poblaciones de comportamiento muy diferente, en donde una de ellas tenía un diagnóstico inicial positivo, mientras que la otra tenía un diagnóstico inicial negativo. Además, se verificó que haciendo una selección según el valor de Estadio FIGO (≤ 4 y ≥ 3) daba unos resultados prácticamente idénticos, simplificando mucho la separación en los grupos de pacientes.

En base a esta separación, se estudiaron las variables de tratamiento utilizando las técnicas de estadística clásica para tratar de determinar cuáles podían afectar más a la supervivencia de la paciente. Para ello, se utilizó el Test exacto de Fisher, que indica si las distribuciones de cada una de las categorías de las variables pueden pertenecer o no a la misma distribución, lo que podría significar que esa variable influyera en la supervivencia de la paciente. Además, para comprobar si esta influencia aumenta o disminuye la supervivencia, se utilizó el intervalo de confianza binomial según la modificación del intervalo de Clopper-Pearson propuesta por Lancaster, que proporciona el intervalo de error del exitus en ese valor de la variable.

Así pues, en el grupo de pacientes con diagnóstico inicial negativo, se observaron algunos indicios de diferencias en la supervivencia según la aplicación de ciertos tratamientos. En especial, se pueden destacar dos variables que tienen indicios de esta influencia. La más importante es el Tipo IQ, que es el tipo de cirugía aplicado a la paciente. En esta variable, se puede observar como aplicando el Tipo IQ 2, que se corresponde a una cirugía de Histerectomía con Doble Anexectomía, parece reducir las posibilidades de supervivencia de la paciente, mientras que si se le aplica a la paciente el Tipo IQ 3, que es un cirugía de Histerectomía con Doble Anexectomía y Linfadenectomía, las posibilidades de supervivencia de la paciente parecen aumentar. Algo parecido ocurre con la variable RTE, que represente el tipo de radioterapia que recibió la paciente. Para este grupo de pacientes, se comprueba como para el valor 0, que significa no haber recibido radioterapia, la supervivencia parece verse afectada negativamente, mientras que aplicando un radioterapia tipo Pelvis + Paraaórtica, el porcentaje de supervivencia de las pacientes parece mejorar. Es importante tener en cuenta que sólo se puede hablar de indicios de estas diferencias, ya que no se puede afirmar nada debido a la cantidad de datos disponibles.

Para el otro grupo de pacientes, se confirma una tendencia que se venía observando desde el comienzo del análisis exploratorio, y es que parece influir positivamente no aplicar a la paciente ningún tratamiento adicional antes que aplicar algún tratamiento. Esto es especialmente claro para algunas variables como Quimioterapia o Carbo Taxol, donde aplicar alguno de estos dos tratamientos

reduce la supervivencia de la paciente. Una posible explicación a este fenómeno puede deberse al hecho de que los datos vienen filtrados, ya que el médico ha tenido que tomar decisiones para el tratamiento de las pacientes. Por lo tanto, es probable que aquellas pacientes que sí se sometieron a un tratamiento fuera porque sufrían de un tumor que sí necesitaba un tratamiento, ya que era más peligroso, mientras que aquellas pacientes que no tuvieron ningún tratamiento fue porque el tumor que tenían no lo necesitaba, ya que no tenía peligro para la paciente. De este modo, los tumores que no tienen ningún tratamiento sería porque realmente no lo necesitan, lo que podría explicar que la supervivencia sea menor en aquellas pacientes que sí han recibido algún tratamiento, debido a la propia gravedad del tumor.

En conclusión, es necesario disponer de más datos para garantizar que estos efectos sean significativos, además de discriminar hasta qué punto son causa o son efecto de todo el proceso médico por el que atravesaron las pacientes. Sin embargo, a pesar de la cantidad de datos disponibles, se han obtenido resultados bastante interesantes que puede merecer la pena investigar con más detalle bajo una supervisión médica.

A. Anexo

A.1. Distribución de las variables

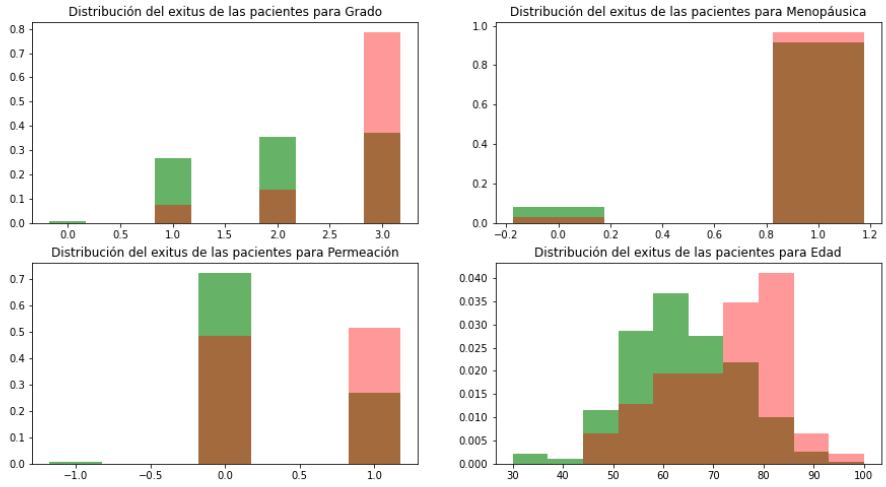

Figura 30: Histogramas y diagramas de barras normalizados de cuatro de las variables del dataset.

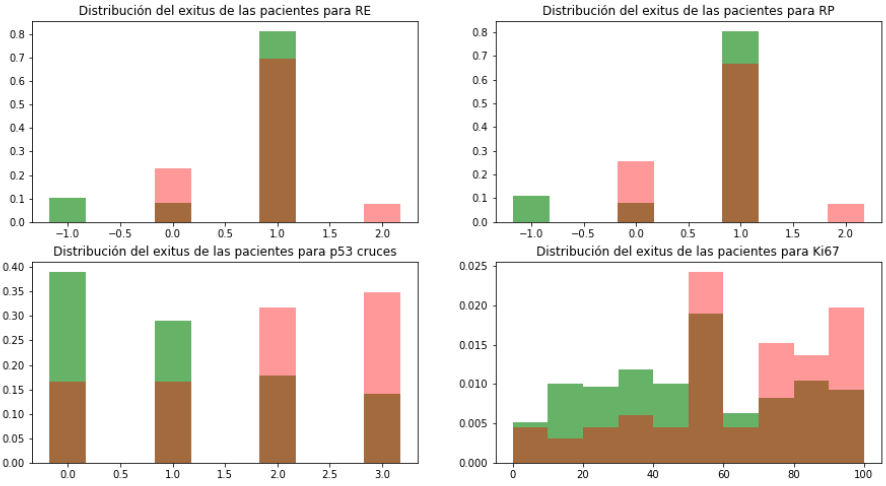

Figura 31: Histogramas y diagramas de barras normalizados de cuatro de las variables del dataset.

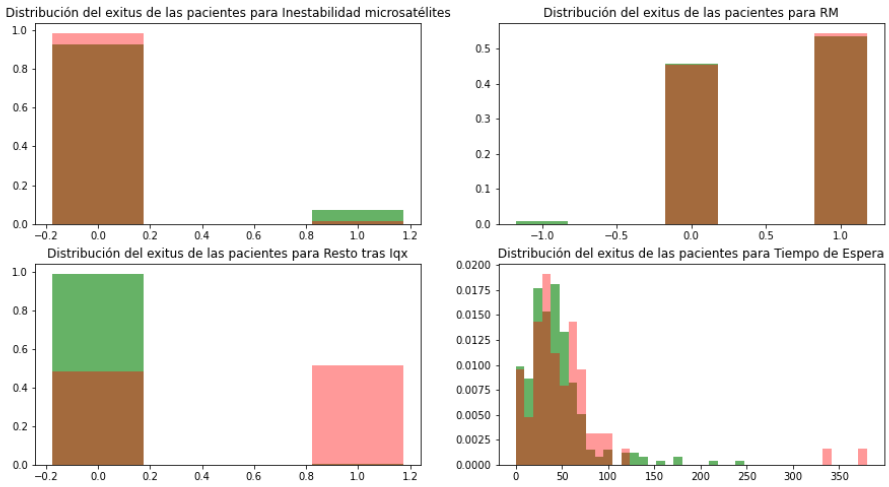

Figura 32: Histogramas y diagramas de barras normalizados de cuatro de las variables del dataset.

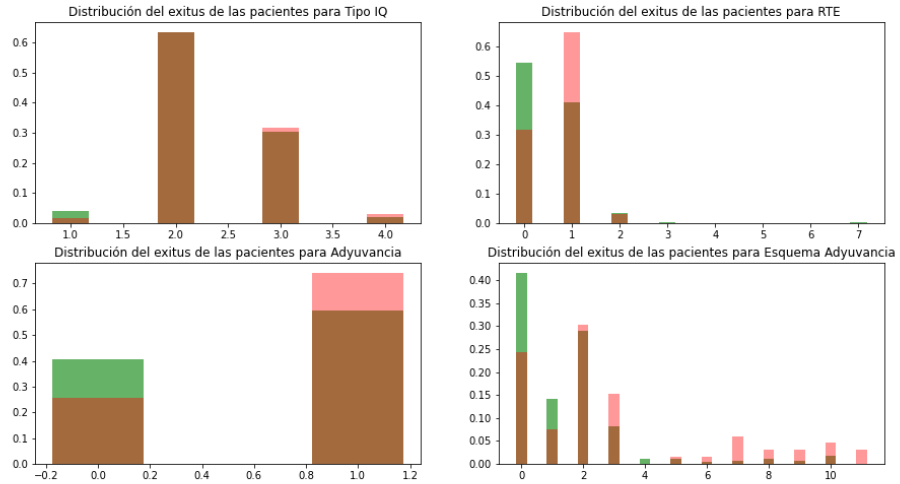

Figura 33: Histogramas y diagramas de barras normalizados de cuatro de las variables del dataset.

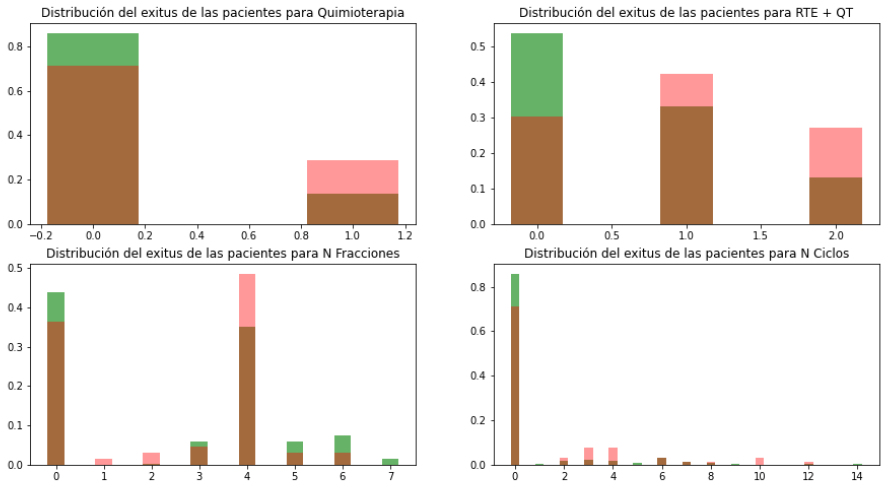

Figura 34: Histogramas y diagramas de barras normalizados de cuatro de las variables del dataset.

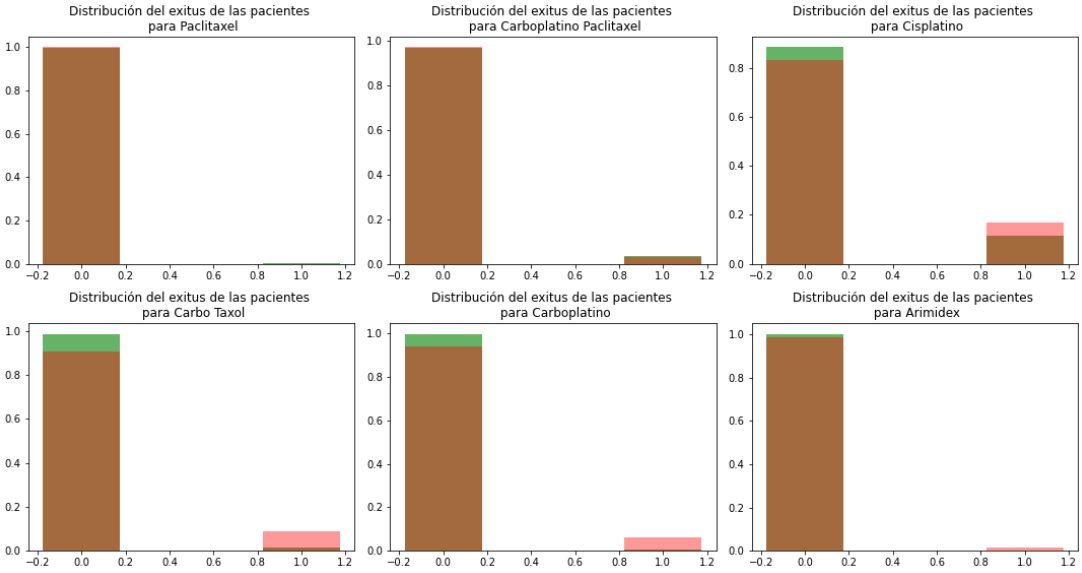

Figura 35: Histogramas y diagramas de barras normalizados de seis de las variables del dataset.

A.2. Distribución de las variables de diagnóstico fijando los valores de Estadio FIGO

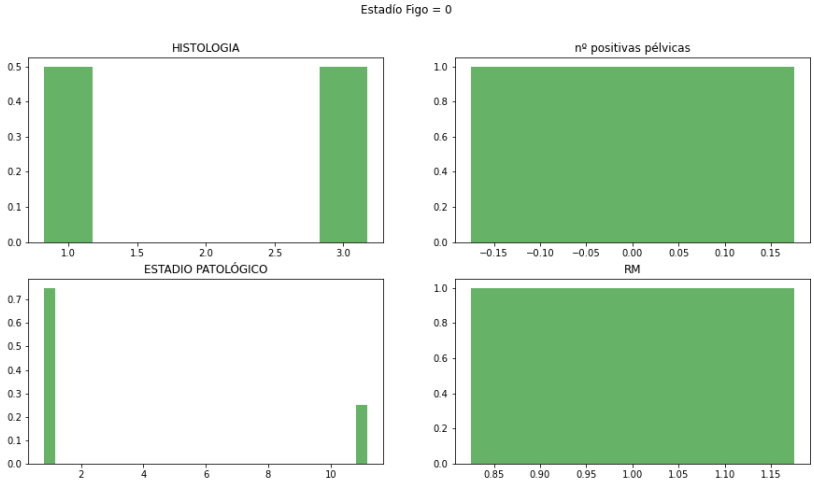

Figura 36: Distribución de las distintas variables de diagnóstico separadas según el exitus para el valor de Estadio FIGO 0.

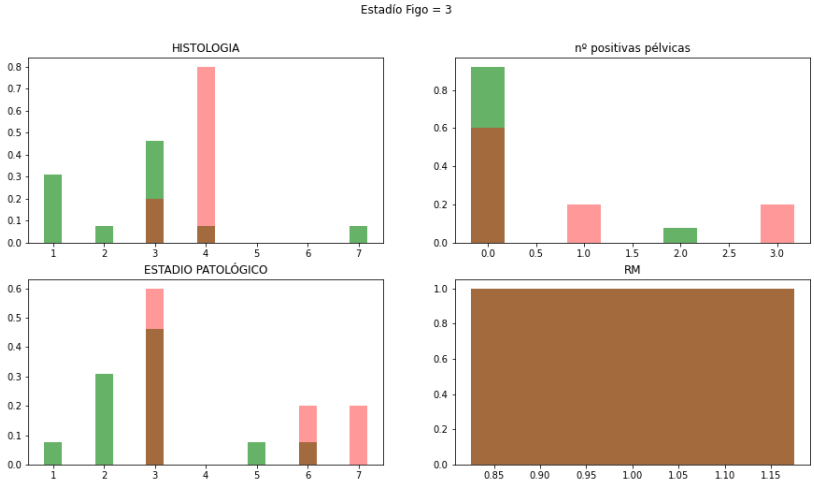

Figura 37: Distribución de las distintas variables de diagnóstico separadas según el exitus para el valor de Estadio FIGO 3.

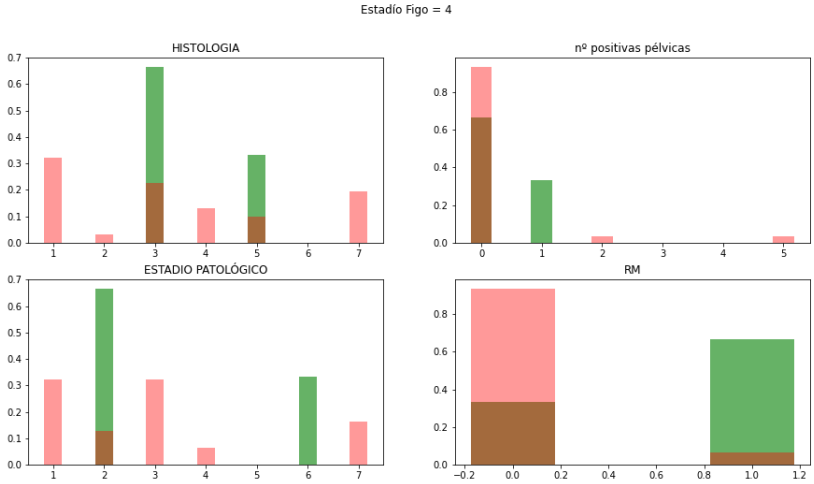

Figura 38: Distribución de las distintas variables de diagnóstico separadas según el exitus para el valor de Estadio FIGO 4.

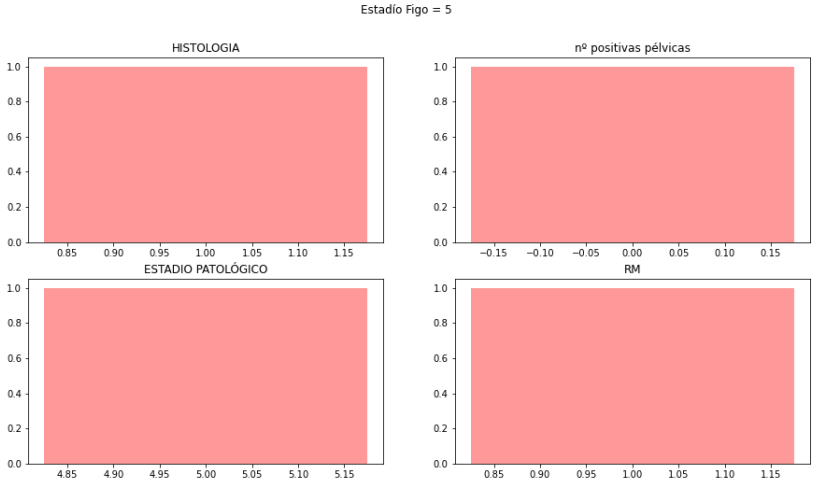

Figura 39: Distribución de las distintas variables de diagnóstico separadas según el exitus para el valor de Estadio FIGO 5.

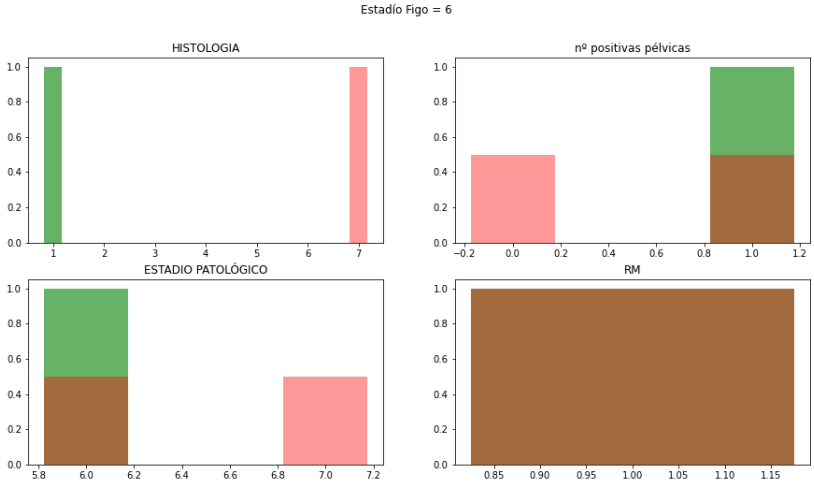

Figura 40: Distribución de las distintas variables de diagnóstico separadas según el exitus para el valor de Estadio FIGO 6.

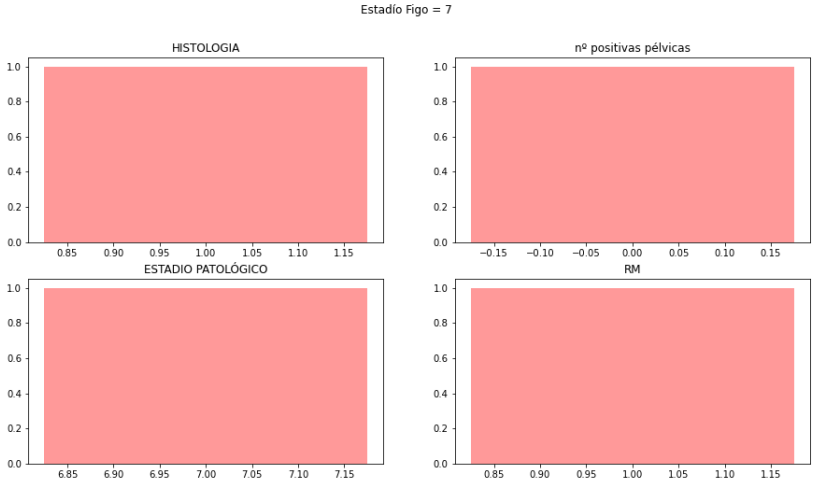

Figura 41: Distribución de las distintas variables de diagnóstico separadas según el exitus para el valor de Estadio FIGO 7.

A.3. Distribución de las variables de tratamiento fijando los valores de Estadio FIGO

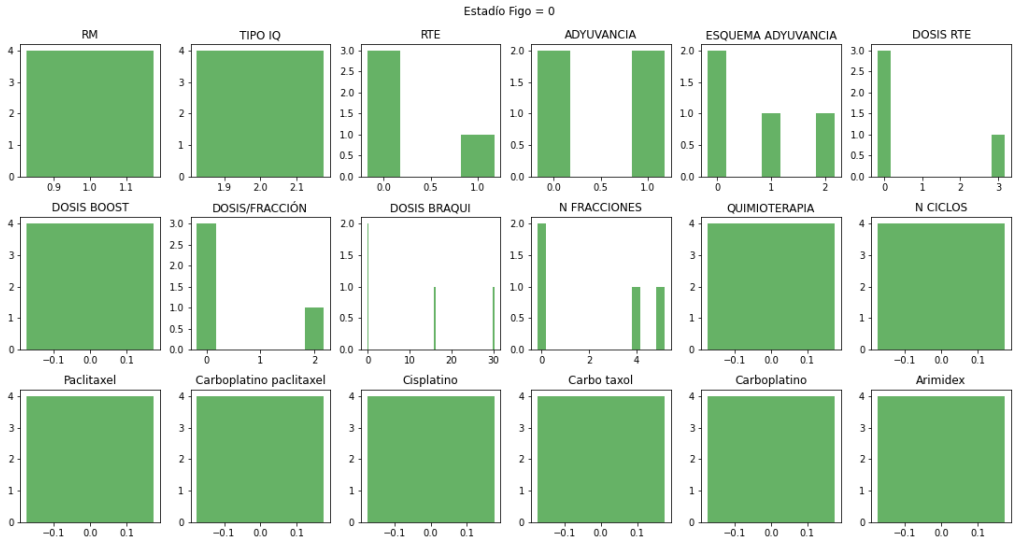

Figura 42: Distribución de las distintas variables de tratamiento separadas según el exitus para el valor de Estadio FIGO 0.

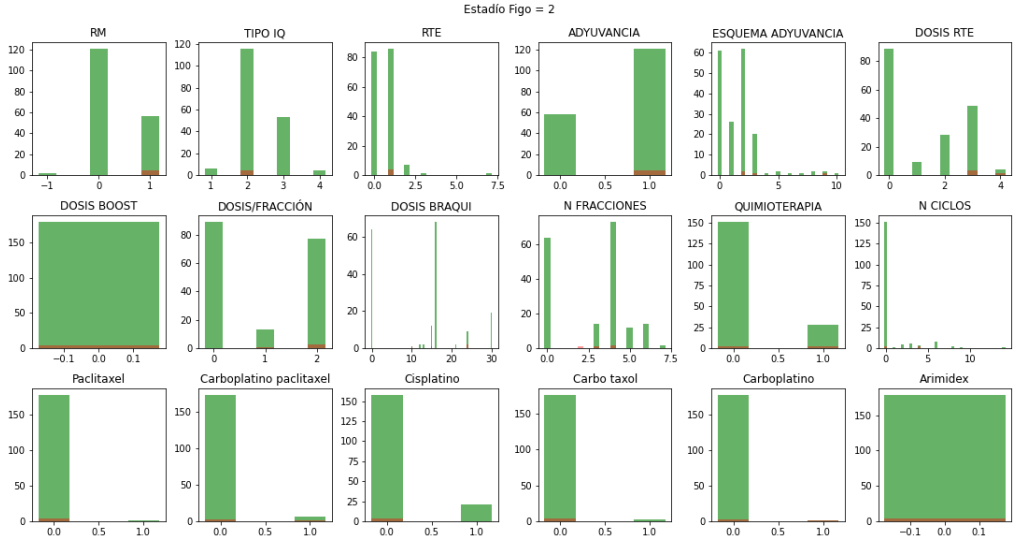

Figura 43: Distribución de las distintas variables de tratamiento separadas según el exitus para el valor de Estadio FIGO 2.

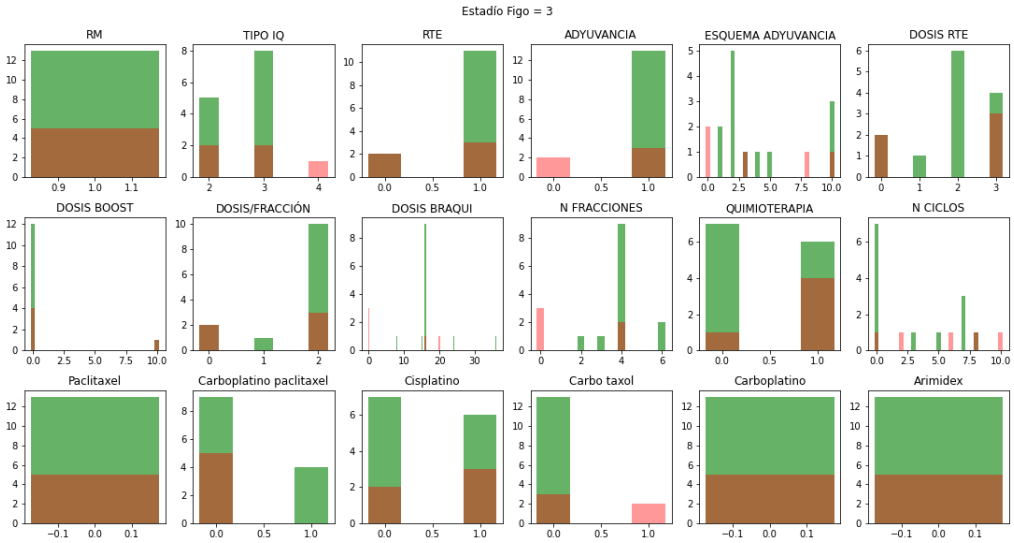

Figura 44: Distribución de las distintas variables de tratamiento separadas según el exitus para el valor de Estadio FIGO 3.

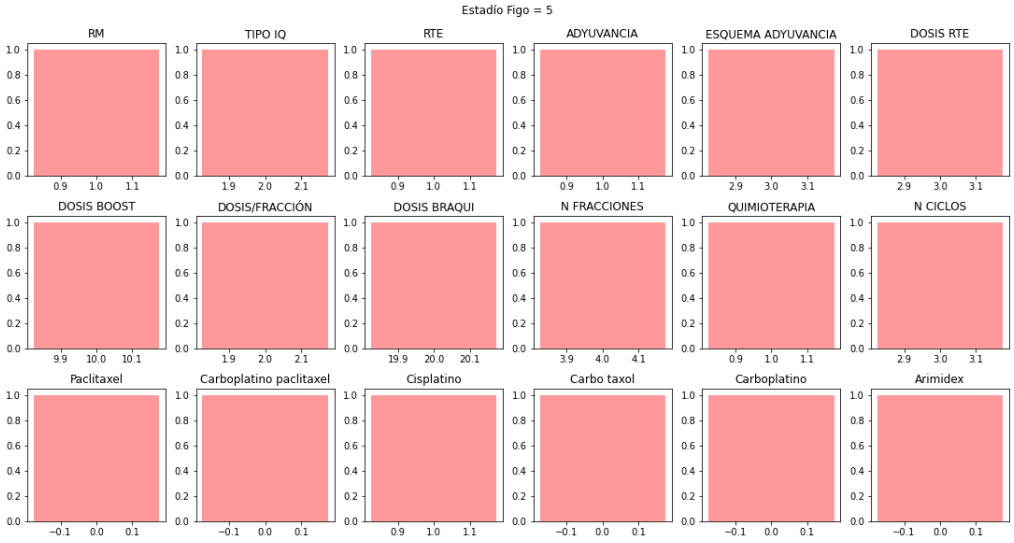

Figura 45: Distribución de las distintas variables de tratamiento separadas según el exitus para el valor de Estadio FIGO 5.

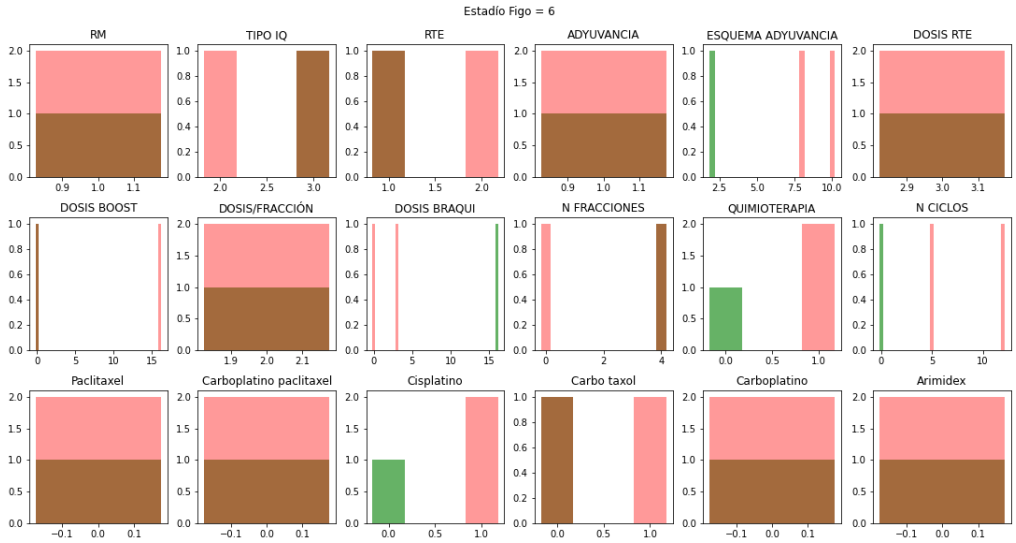

Figura 46: Distribución de las distintas variables de tratamiento separadas según el exitus para el valor de Estadio FIGO 6.

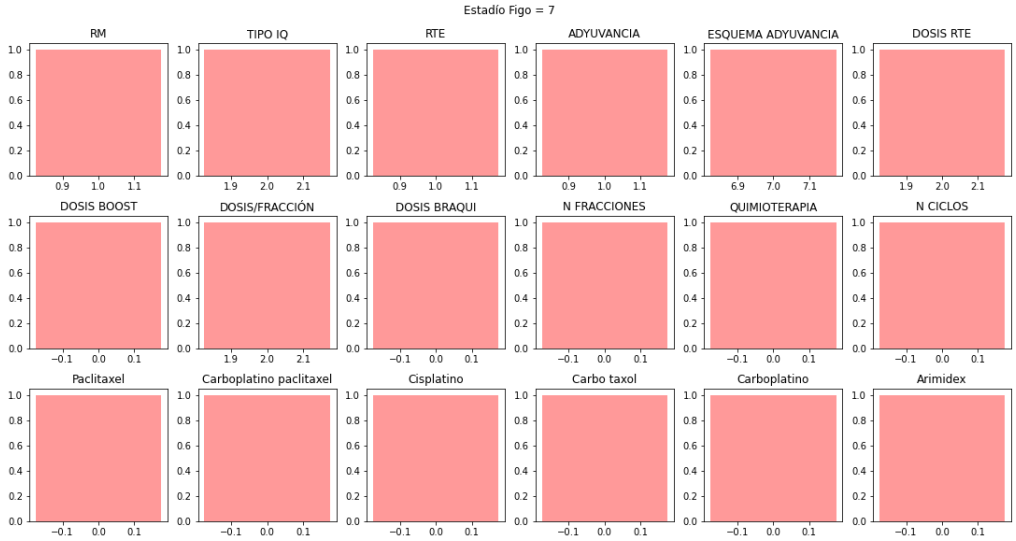

Figura 47: Distribución de las distintas variables de tratamiento separadas según el exitus para el valor de Estadio FIGO 7.

A.4. Distribución de las variables de tratamiento en función de la salida del árbol entrenado con las variables de diagnóstico

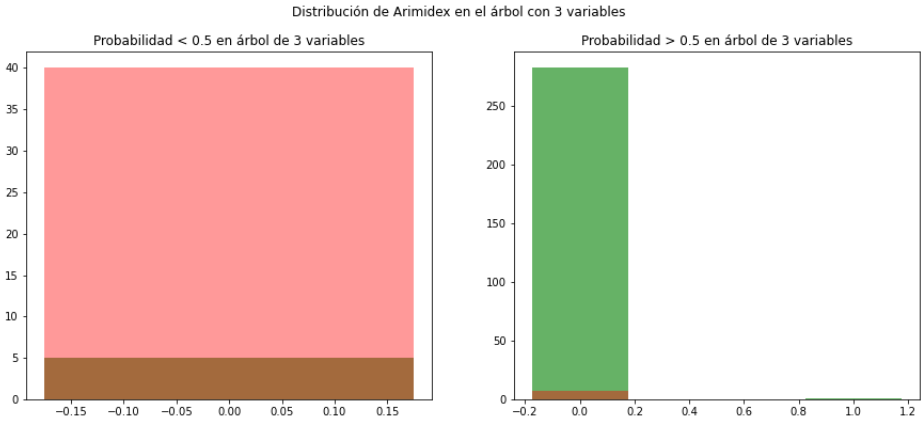

Figura 48: Distribución de Arimidex dividida según la predicción del árbol de las variables de diagnóstico.

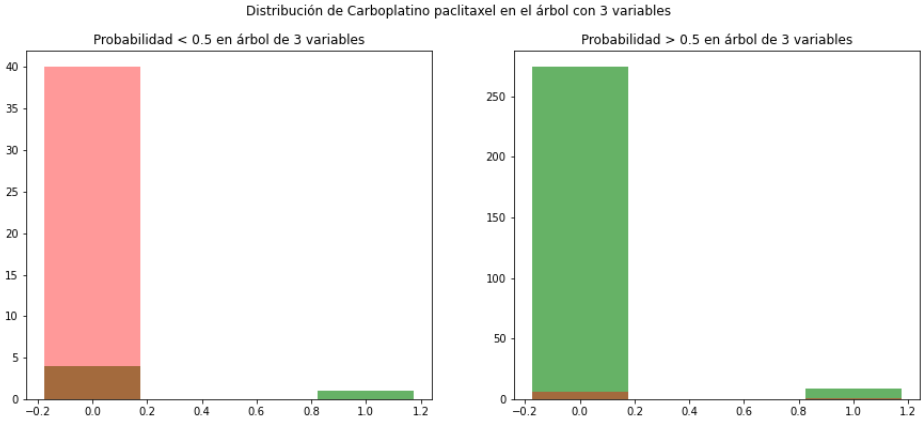

Figura 49: Distribución de Carboplatino Paclitaxel dividida según la predicción del árbol de las variables de diagnóstico.

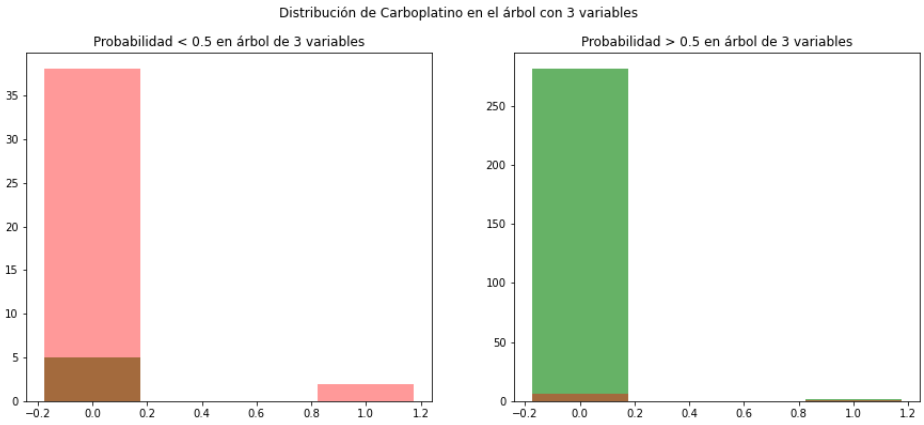

Figura 50: Distribución de Carboplatino dividida según la predicción del árbol de las variables de diagnóstico.

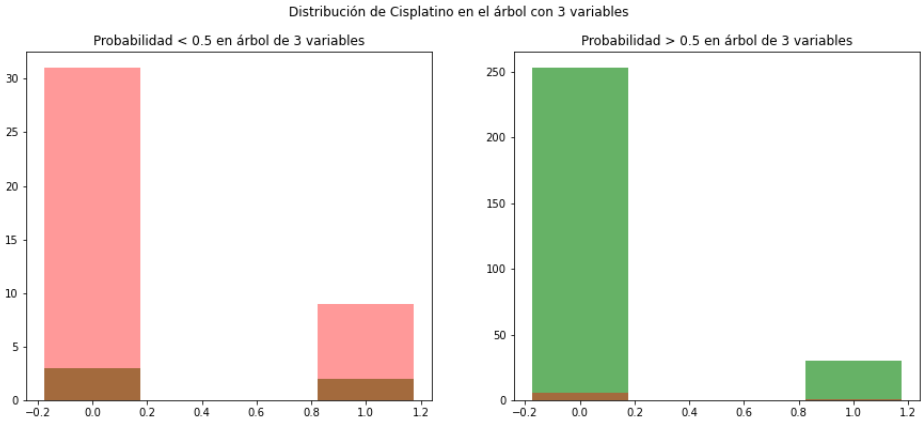

Figura 51: Distribución de Cisplatino dividida según la predicción del árbol de las variables de diagnóstico.

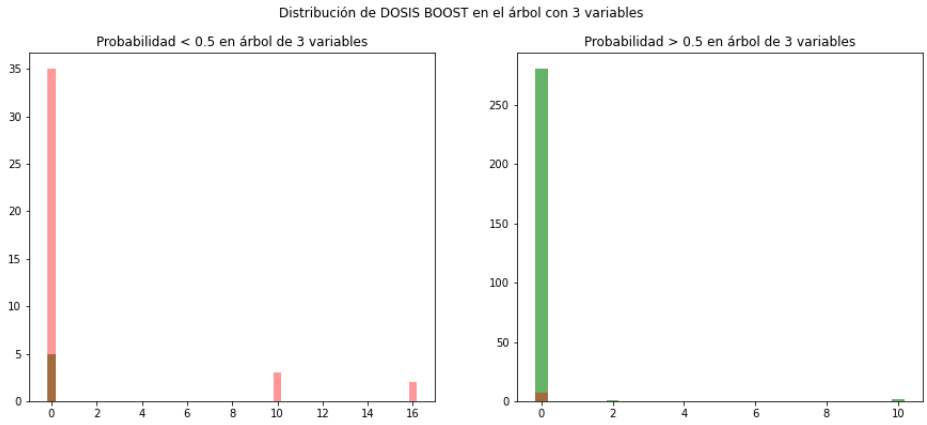

Figura 52: Distribución de Dosis Boost dividida según la predicción del árbol de las variables de diagnóstico.

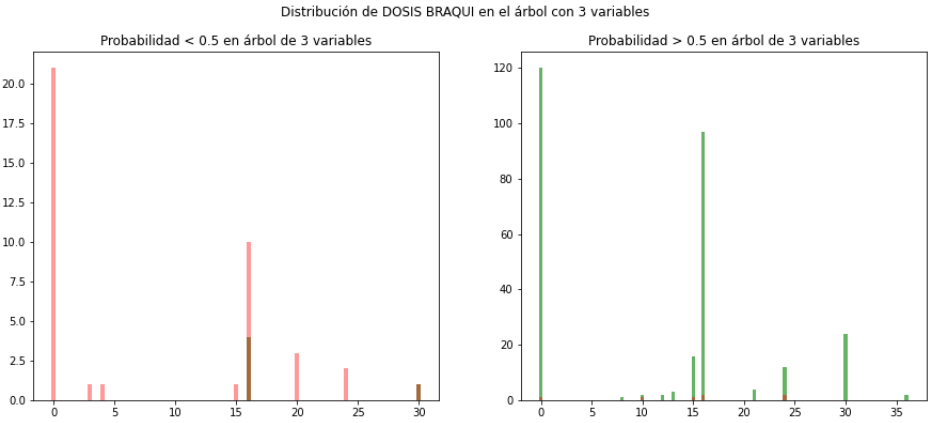

Figura 53: Distribución de Dosis Braqui dividida según la predicción del árbol de las variables de diagnóstico.

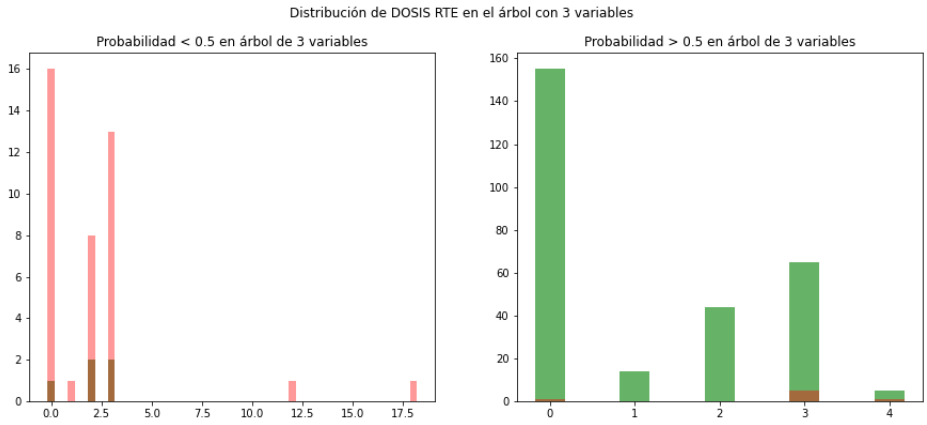

Figura 54: Distribución de Dosis RTE dividida según la predicción del árbol de las variables de diagnóstico.

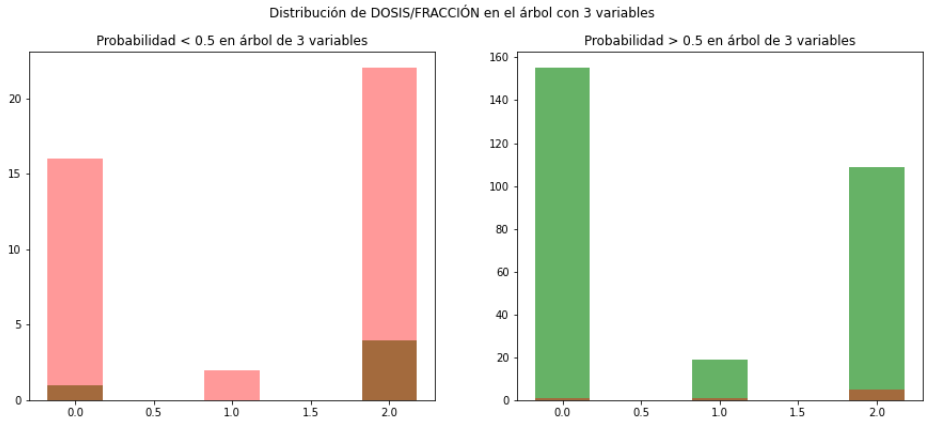

Figura 55: Distribución de Dosis Fracción dividida según la predicción del árbol de las variables de diagnóstico.

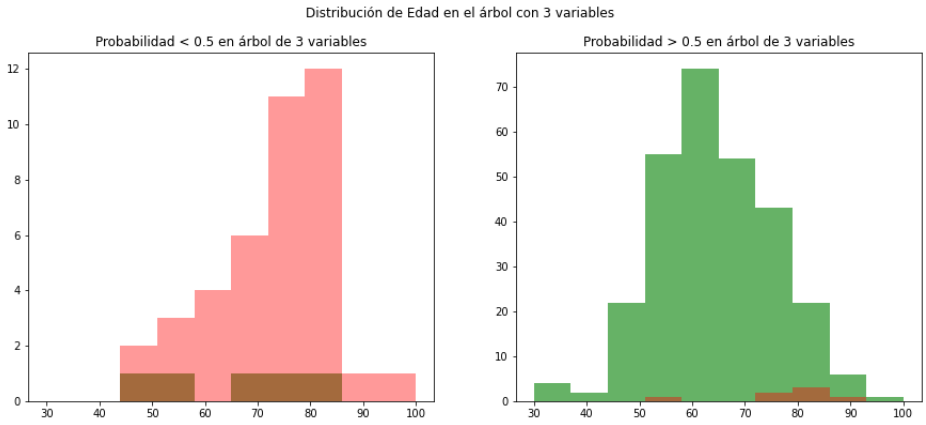

Figura 56: Distribución de Edad dividida según la predicción del árbol de las variables de diagnóstico.

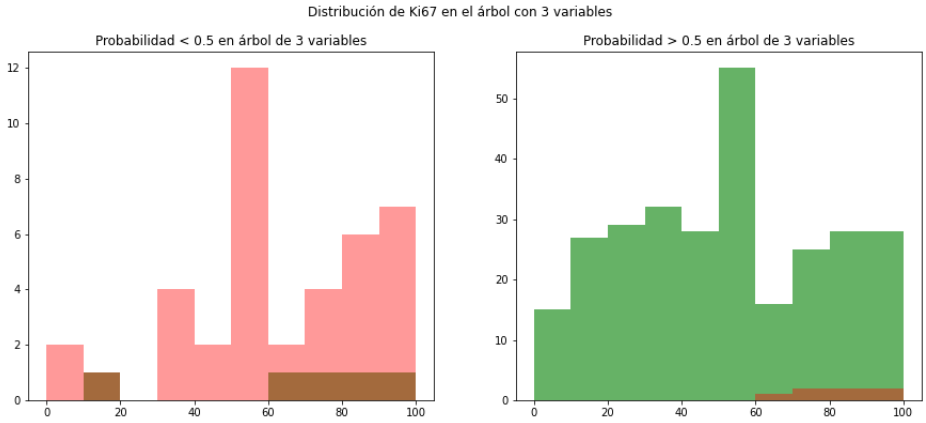

Figura 57: Distribución de Ki67 dividida según la predicción del árbol de las variables de diagnóstico.

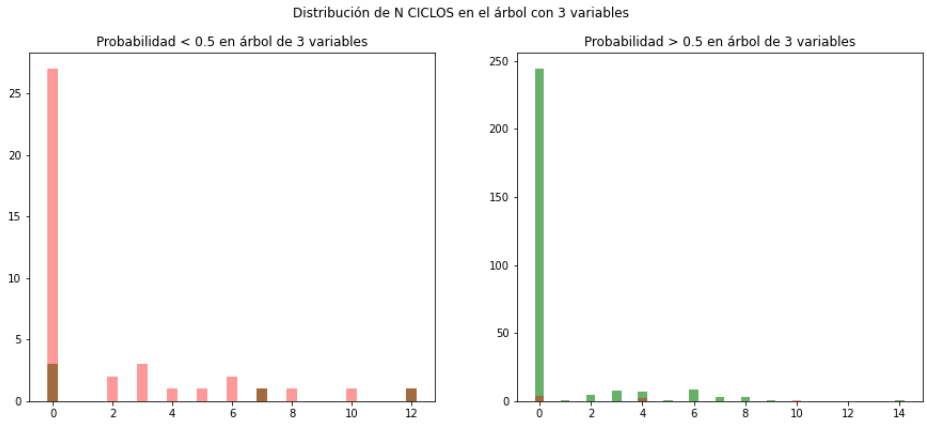

Figura 58: Distribución de N Ciclos dividida según la predicción del árbol de las variables de diagnóstico.

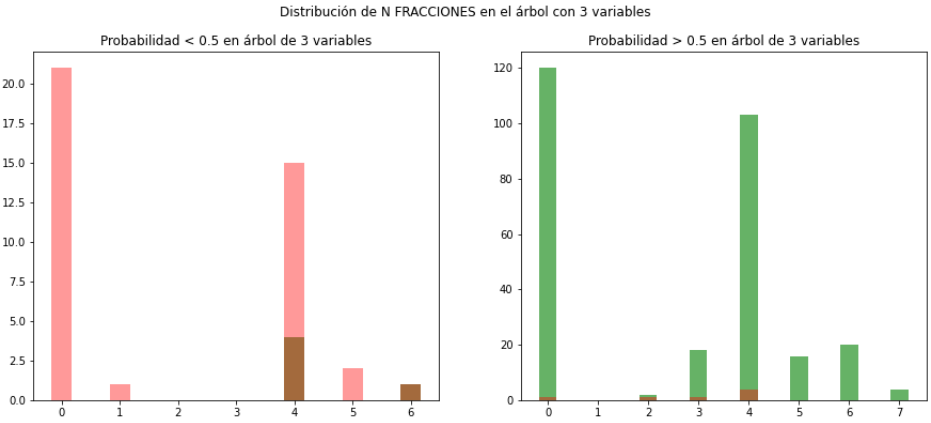

Figura 59: Distribución de N Fracciones dividida según la predicción del árbol de las variables de diagnóstico.

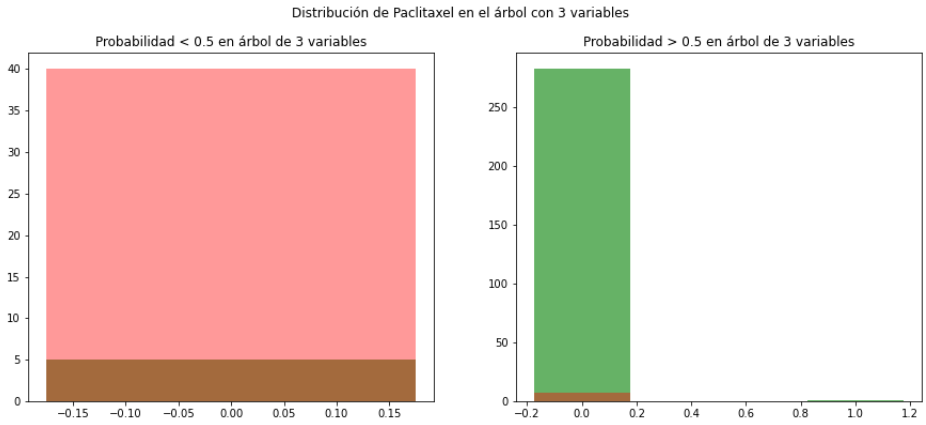

Figura 60: Distribución de Paclitaxel dividida según la predicción del árbol de las variables de diagnóstico.

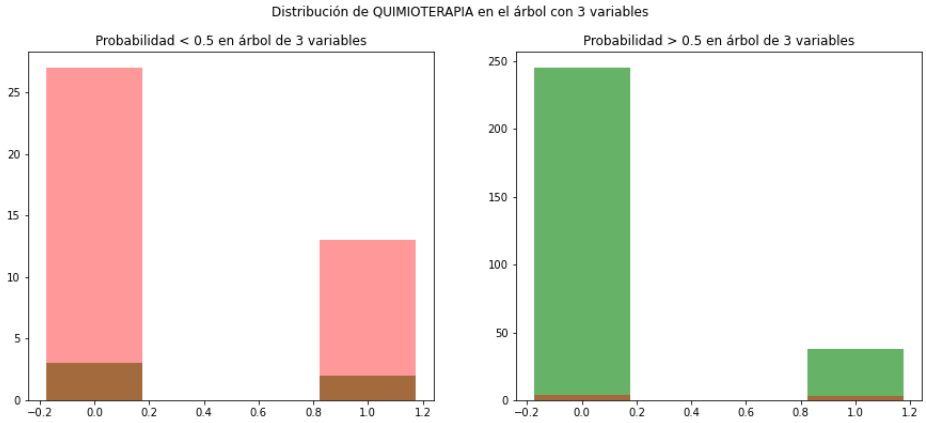

Figura 61: Distribución de Quimioterapia dividida según la predicción del árbol de las variables de diagnóstico.

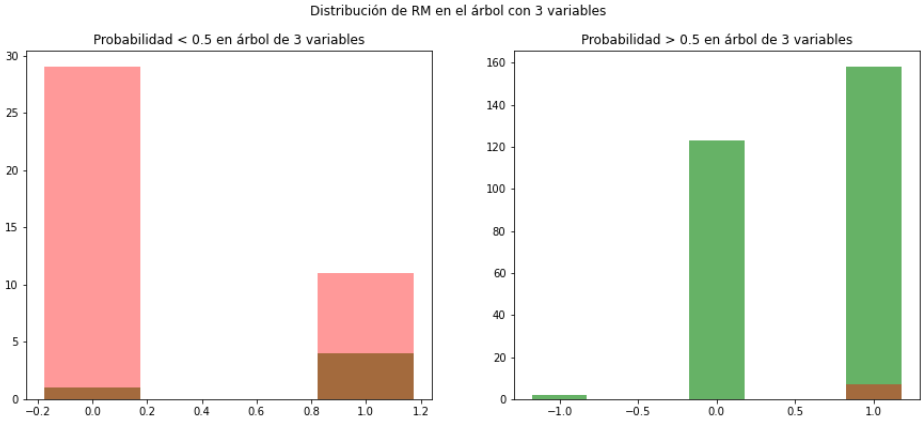

Figura 62: Distribución de Resonancia Magnética dividida según la predicción del árbol de las variables de diagnóstico.

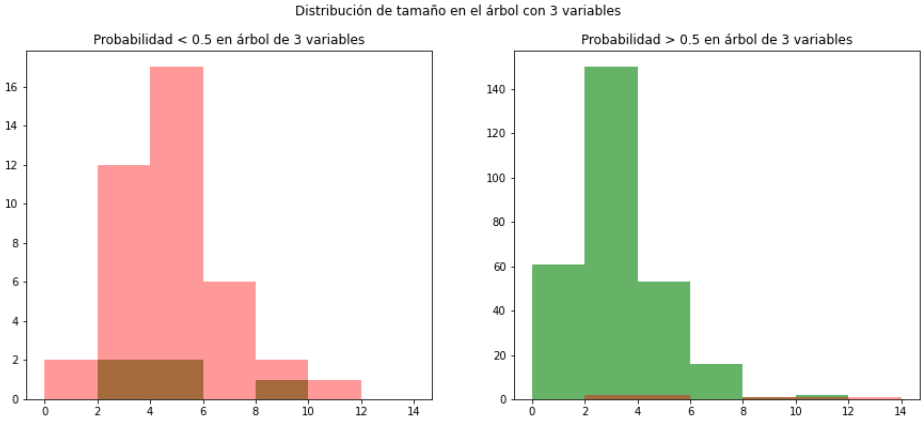

Figura 63: Distribución de Tamaño dividida según la predicción del árbol de las variables de diagnóstico.

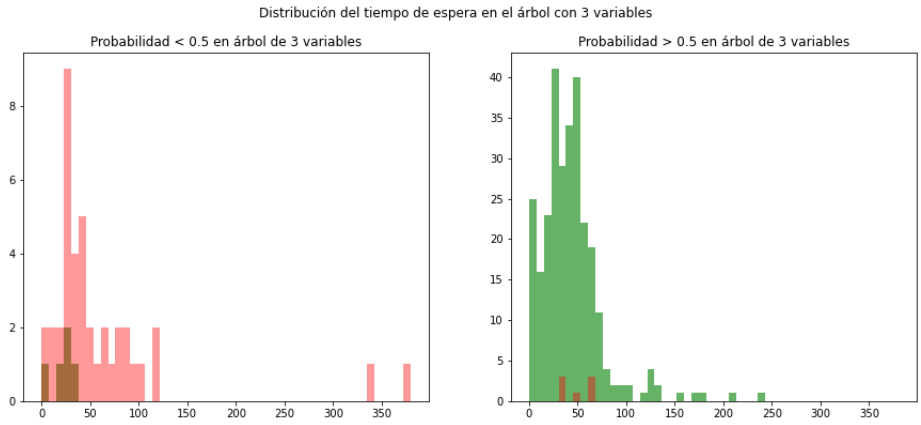

Figura 64: Distribución de Tiempo de Espera dividida según la predicción del árbol de las variables de diagnóstico.

A.5. Distribución de las variables de tratamiento en función del valor de Estadio FIGO

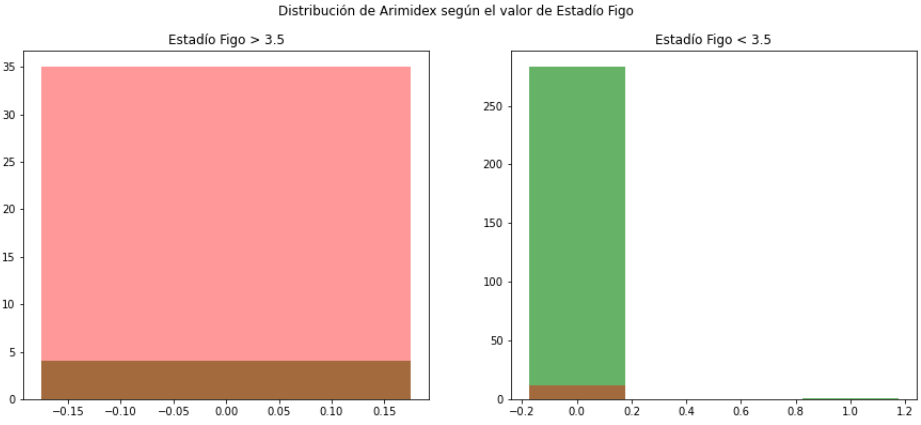

Figura 65: Distribución de Arimidex dividida según el valor de Estadio FIGO.

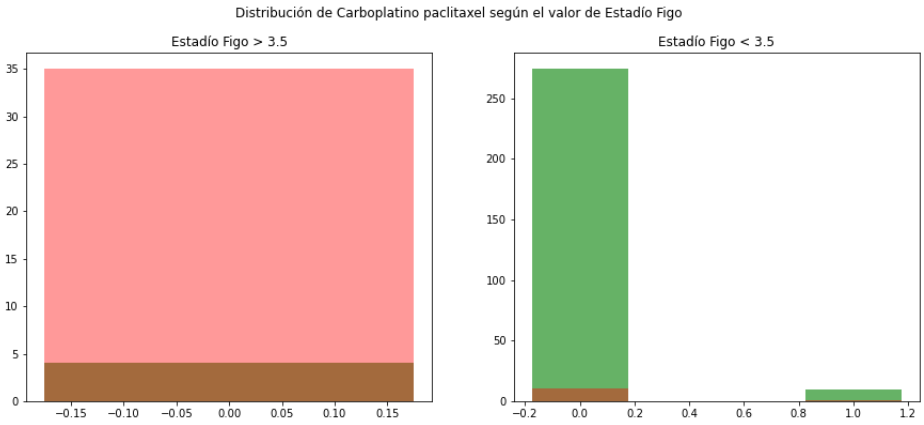

Figura 66: Distribución de Carboplatino Paclitaxel dividida según el valor de Estadio FIGO.

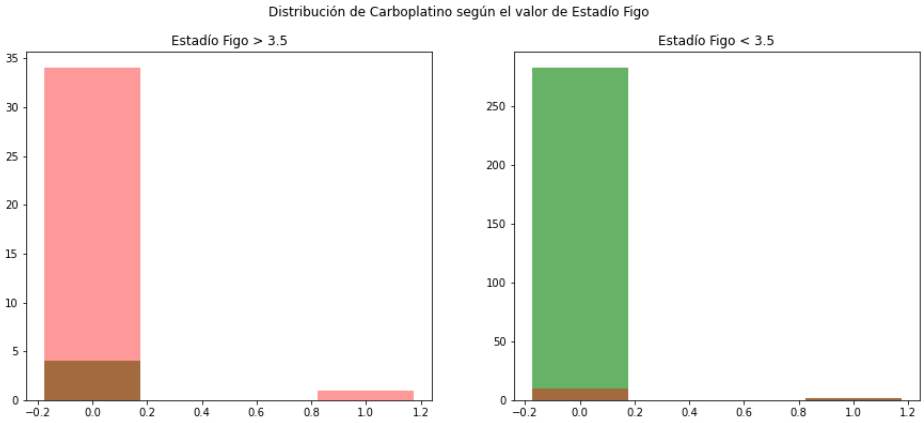

Figura 67: Distribución de Carboplatino dividida según el valor de Estadio FIGO.

Figura 68: Distribución de Cisplatino dividida según el valor de Estadio FIGO.

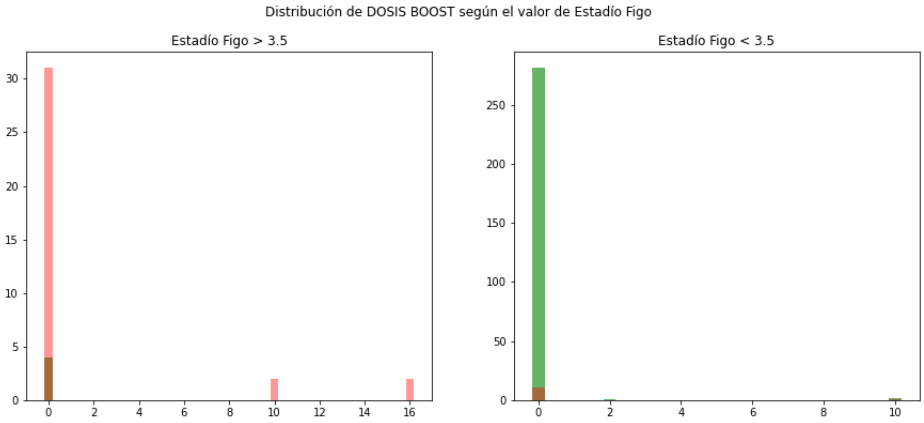

Figura 69: Distribución de Dosis Boost dividida según el valor de Estadio FIGO.

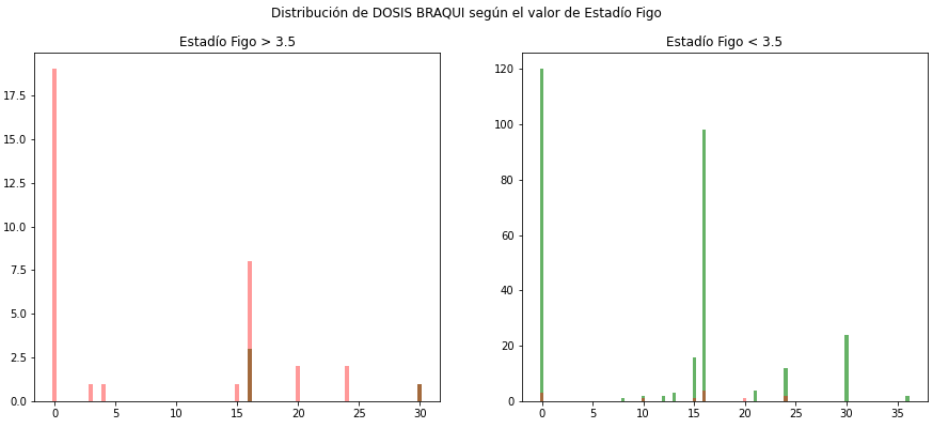

Figura 70: Distribución de Dosis Braqui dividida según el valor de Estadio FIGO.

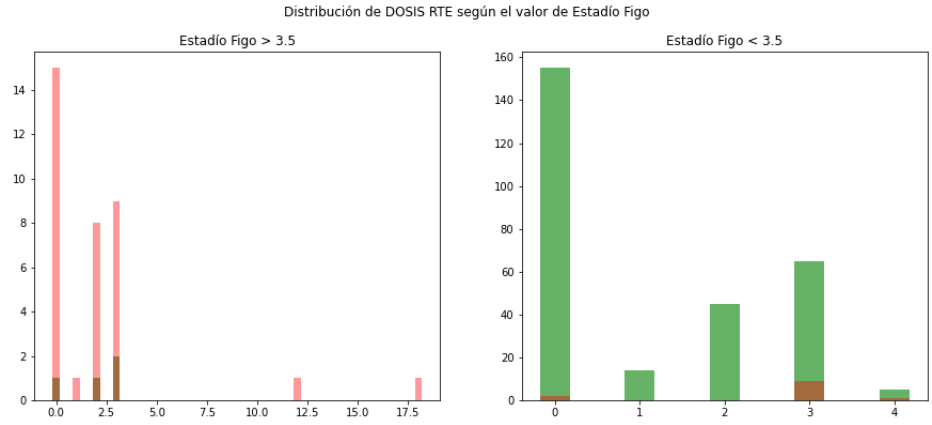

Figura 71: Distribución de Dosis RTE dividida según el valor de Estadio FIGO.

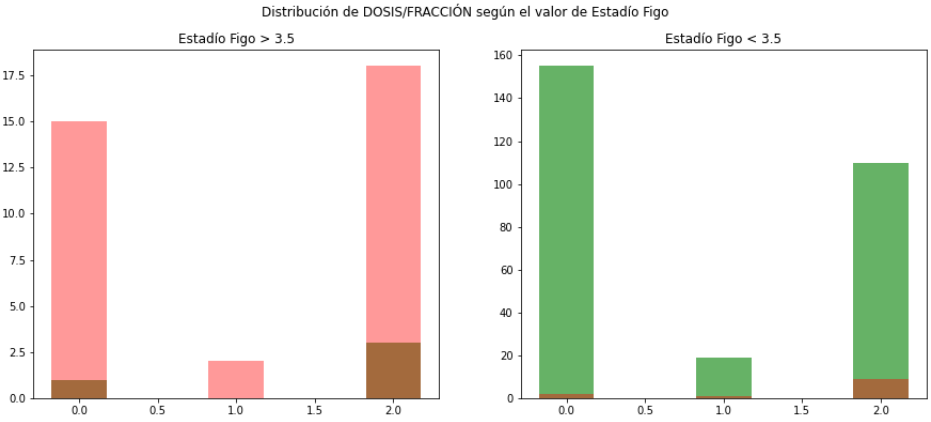

Figura 72: Distribución de Dosis Fracción dividida según el valor de Estadio FIGO.

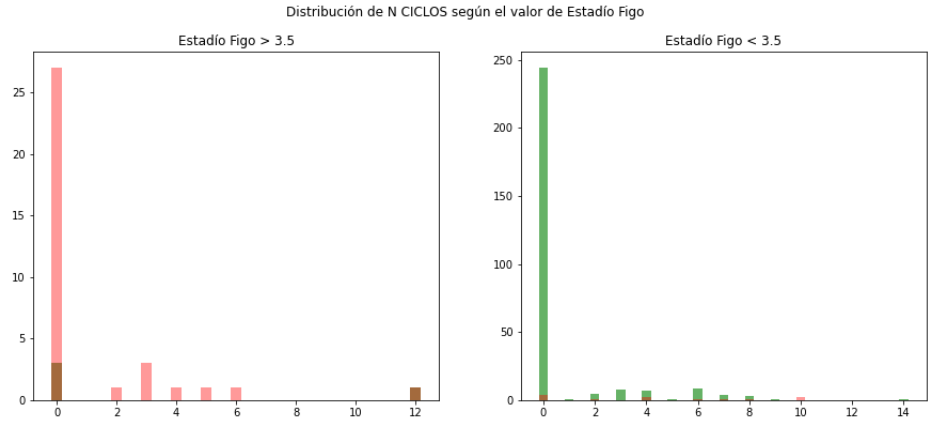

Figura 73: Distribución de N Ciclos dividida según el valor de Estadio FIGO.

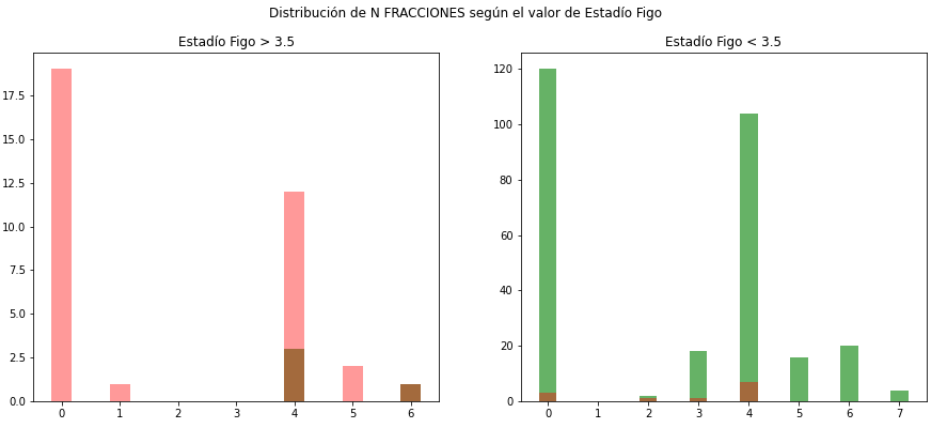

Figura 74: Distribución de N Fracciones dividida según el valor de Estadio FIGO.

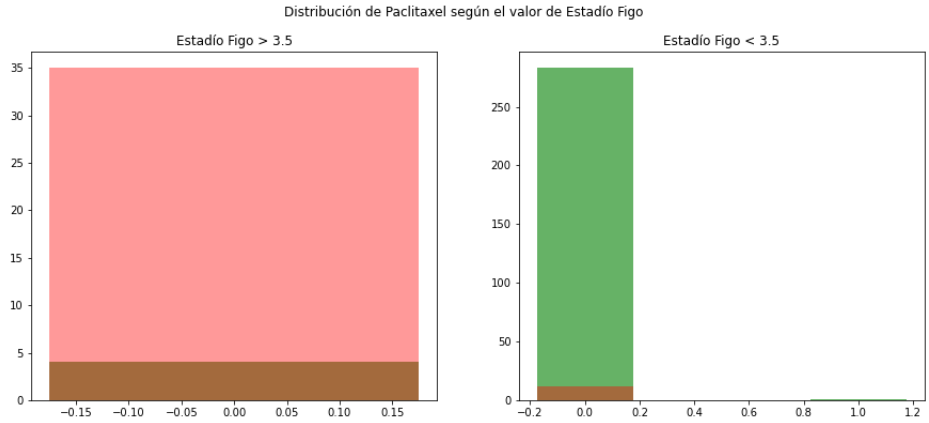

Figura 75: Distribución de Paclitaxel dividida según el valor de Estadio FIGO.

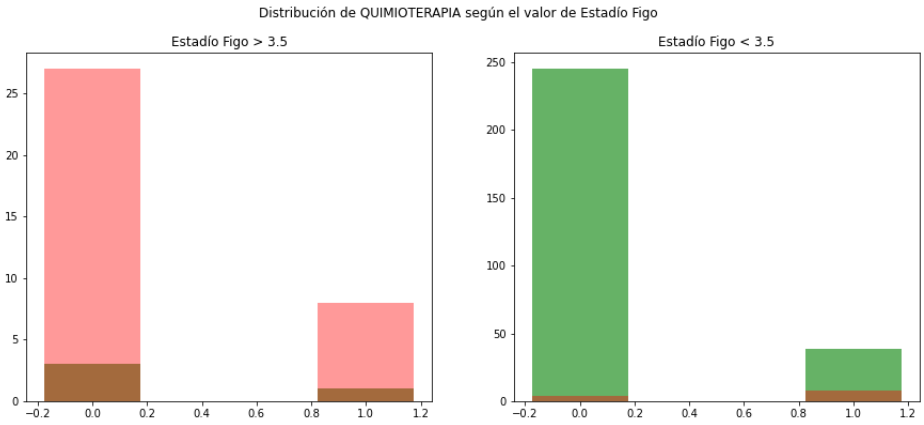

Figura 76: Distribución de Quimioterapia dividida según el valor de Estadio FIGO.

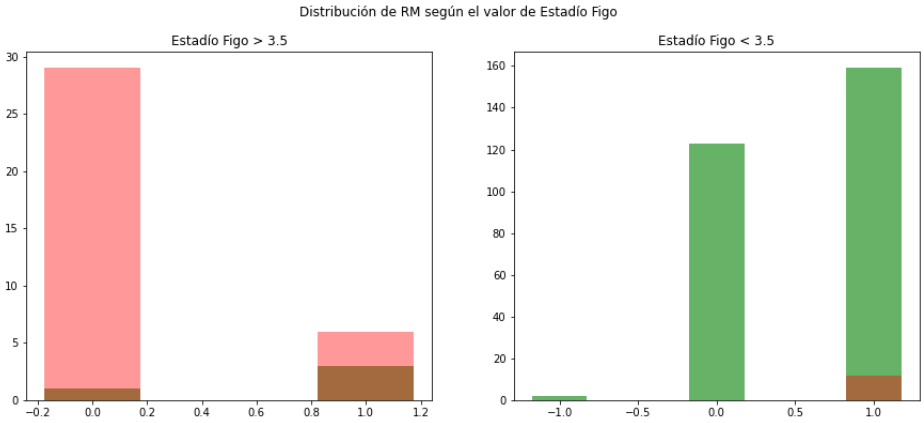

Figura 77: Distribución de Resonancia Magnética dividida según el valor de Estadio FIGO.

Referencias

[1] Guido Van Rossum y Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[2] The pandas development team. *pandas-dev/pandas: Pandas*. Ver. latest. Feb. de 2020. DOI: 10.5281/zenodo.3509134.

[3] Wes McKinney. *Data Structures for Statistical Computing in Python*. Ed. por Stéfan van der Walt y Jarrod Millman. 2010, págs. 56-61. DOI: 10.25080/Majora-92bf1922-00a.

[4] Robin Genuer y Jean-Michel Poggi. "Random Forests". En: sep. de 2020, págs. 33-55. DOI: 10.1007/978-3-030-56485-8_3.

[5] Brad Boehmke y Brandon Greenwell. "Gradient Boosting". En: nov. de 2019, págs. 221-246. ISBN: 9780367816377. DOI: 10.1201/9780367816377-12.

[6] Fabian et al. Pedregosa. "Scikit-learn: Machine Learning in Python". En: *Journal of Machine Learning Research* 12 (ene. de 2012).

[7] RA Fisher. *Statistical methods for research workers*. Oliver & Boyd, Edinburgh, 1934.

[8] C. J. Clopper y E. S. Pearson. "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial". En: *Biometrika* 26.4 (1934), págs. 404-413.

[9] H. O. Lancaster. "Significance Tests in Discrete Distributions". En: *Journal of the American Statistical Association* 56.294 (1961), págs. 223-234.